DATA SCIENCE APPROACH FOR SPATIOTEMPORAL MODELING OF DENGUE IN PUNJAB, INDIA

Dr Gurpreet Singh

PhD THESIS

2023

SREE CHITRA TIRUNAL INSTITUTE FOR MEDICAL SCIENCES AND TECHNOLOGY, TRIVANDRUM

An Institution of National Importance,

Dept. of Science and Technology, Govt. of India

www.sctimst.ac.in

TITLE OF THESIS

DATA SCIENCE APPROACH FOR SPATIOTEMPORAL MODELING OF DENGUE IN PUNJAB, INDIA

A THESIS SUBMITTED BY

Dr Gurpreet Singh

TO

SREE CHITRA TIRUNAL INSTITUTE FOR MEDICAL SCIENCES AND TECHNOLOGY, TRIVANDRUM.

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE AWARD OF

DOCTOR OF PHILOSOPHY

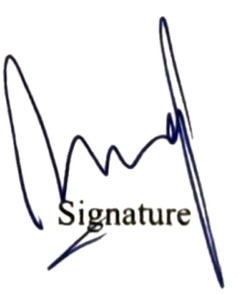
2023

DECLARATION BY THE STUDENT

CERTIFICATE

I, Dr Gurpreet Singh hereby certify that I had personally carried out the work depicted in the thesis titled, "Data Science approach for Spatiotemporal Modeling of Dengue in Punjab, India"

No part of this thesis has been submitted for the award of any other degree or diploma prior to this date.



Signature

Dr Gurpreet Singh

Date 26/07/23

CERTIFICATE BY THE RESEARCH GUIDE

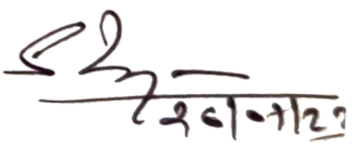
Name of the Guide: Dr Biju Soman

Division/Department: Achutha Menon Centre for Health Science Studies

This is to certify that (name of the student) Dr Gurpreet Singh, department/division of Achutha Menon Centre for Health Science Studies of this institute has fulfilled the requirements prescribed for the Ph.D. degree of the Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum.

The thesis entitled, "Data Science approach for Spatiotemporal Modeling of Dengue in Punjab, India" was carried out under my direct supervision. No part of the thesis was submitted for the award of any degree or diploma prior to this date.

Clearance was obtained from the Institutional Ethics Committee for carrying out the study.

Signature 

Dr Biju Soman

Date 26/07/2023

APPROVAL OF THE THESIS

The thesis entitled

"Data Science approach for Spatiotemporal Modeling of Dengue in Punjab, India"

Submitted by

Dr Gurpreet Singh

for the degree of

Doctor of Philosophy

of

SREE CHITRA TIRUNAL INSTITUTE FOR MEDICAL SCIENCES AND TECHNOLOGY, TRIVANDRUM

is evaluated and approved by

2023-07-26

.....

(Name & Signature of the Guide)

Dr Biju Soman,
Professor & Head
AMCHSS, SCTIMST, Trivandrum.

.....

DR. ARUN KR. SHARMA

(Name & Signature of thesis examiner)

Dr. Arun Kumar Sharma
MD, DNBE, MBA
Professor, Dept. of Community Medicine
University College of Medical Sciences
(University of Delhi),
Dilshad Garden, Delhi-110095

ACKNOWLEDGEMENTS

This thesis has represented another marvelous chapter in my life and is now complete. I acknowledge my heartfelt thanks and gratitude to the lord almighty, the creator of life whose blessings paved the path toward success throughout the journey.

I want to thank the Ministry of Defence, who gave me an opportunity to take study leave to pursue PhD and Sree Chitra Tirunal Institute for Medical Science and Technology for allowing me to undertake the PhD program in an institution of National Importance.

The journey would never have been full of learning without guidance and support from my guide. I want to place on record my heartfelt thanks and gratitude to Dr Biju Soman, my supervisor and guide, who helped me develop skills and undertake this research. He has not only guided me unconditionally throughout the academic journey but has been a mentor for the learnings of life. I want to thank him for his support and guidance throughout this journey.

I would also like to thank my Doctoral Advisory Committee: Dr Jeemon P, Dr Srikant A, Dr Manojkumar TK, and Dr Shijulal Nelson Sathi, who provided constant support and guidance during the research journey. I am thankful to the Doctor Advisory Committee for their constructive advisories and timely corrections.

I extend my heartiest thanks to the faculty members at Achutha Menon Centre for Health Science Studies for their critical comments and counsel in times of need. I place

my utmost appreciation to Dr Arun Mitra, Dr Adrija Roy, Dr Antony, Dr Bevin, and all other colleagues for their valuable encouragement and contribution to my research process.

I place on record my thanks and gratefulness to the Registrar, Deputy Registrar, administrative staff at the Division of Academic Affairs, and team at the Computer Division for their wholehearted support.

I stand grateful to my parents, Satpal Singh and Raj Kaur, who have laid the foundation for my academic journey. I am ever thankful for their sacrificial help and prayers throughout this research. I am also indebted to my sister, Jaspreet Kaur and her husband, Gurpreet Singh, for their ever-present support and encouragement.

I extend my heartful appreciation to my parents-in-law, Hardeep Singh and Baljit Kaur, for their tireless support and prayers all along this journey.

I place on record my deepest gratitude and indebtedness to my wife, Dr Prabhjot Kaur, for her countless sacrifices and for walking along this journey cheerfully.

I am honored to enlist my gratefulness to my daughter, Anahad Kaur, and my dogs, Zorro, Zubin, and Zoozoo, who have been my joy and poured gladness into my heart as they witnessed the journey and struggles at the office and home.

Finally, I am grateful to all my friends, colleagues, faculty, family, and relatives for their encouragement throughout this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

S.No.	Abbreviation	Full Form
1	AIC	Akaike Information Criteria
2	API	Application Programming Interface
3	ARIMA	Autoregressive Integrated Moving Averages
4	CNA	Citation Network Analysis
5	DAAC	Distributed Active Archive Centers
6	DENV	Dengue Virus
7	DTR	Diurnal Temperature Ranges
8	EDA	Exploratory Data Anaalysis
9	EIP	External Incubation Period
10	GHSL	Global Human Settlement Layer
11	GIS	Geographical Information System
12	GoI	Government of India
13	HDF	Hierarchial Data Format
14	IDSP	Integrated Disease Surveillance Programme
15	IHIP	Integrated Health Information Portal
16	IMD	Indian Meteorological Department
17	IMERG	Integrated Multi-satellitE Retrievals of the Global Precipitation Mission
18	LMICs	Low- and Middle- Income Countries
19	MERRA	Modern Era Retrospective-Analysis for Research and Applications

20	MODIS	Moderate Resolution Imaging Spectroradiometer
21	netCDF	network Common Data Form
22	NVBDCP	National Vector Borne Disease Control Programme
23	OCM2	Ocean Color Monitor Version 2
24	POWER	Prediction of Worldwide Energy Resources
25	GAM	Generalized Additive Model
26	GAMM	Generalized Additive Mixed Model
27	GLM	Generalized Linear Models
28	GMAO	Global Modeling and Assimilation Office
29	NDVI	Normalized Difference Vegetation Index
30	PCR	Polymerase Chain Reaction
31	QQ plot	Quantile-quantile plot
32	RHIS	Routine Health Information Systems
33	RMSE	Root Mean Squared Error
34	SEAR	South-East Asian Region
35	SIR	Standardized Incidence Ratios
36	sq. km	Square Kilometres
37	WoS	Web of Science
38	WHO	World Health Organization

SYNOPSIS

Dengue is a significant public health problem in India. Routine Health Information Systems (RHIS) data is rich in information and often used for administrative purposes and as a public health decision-support tool. Technological advancements in Geographical Information Systems (GIS) and data analytics are recommended for strengthening surveillance mechanisms. The emergence of Data Science, an interdisciplinary discipline, provides solutions and strategies for managing routine datasets. However, its potential for extracting valuable knowledge from RHIS remains poorly explored in India.

Methods. The present study explored mechanisms that generate knowledge from RHIS data in resource-constrained settings. We undertook the secondary data analysis of routine health data and linked it with open data sources from the non-health sectors. The study design was an ecological study using a data science approach. The data science approach included data-driven modeling through an iterative data exploration and analysis. The major phases of the study were data collection, extraction, and pre-processing; Exploratory Data Analysis (EDA); Data analysis and interpretation; and development and evaluation of Dengue forecasting models.

Data Sources. We collected line-listing data of lab-confirmed Dengue cases reported by the National Vector Borne Disease Control Programme (NVBDCP), Punjab, from 01 Jan 2015- 31 Dec 2019. We also collected data on climatic and environmental risk factors from satellite imagery datasets and socio-demographic factors from the Census of India 2011 data tables. Population projections were

calculated, and data on urbanization and built-up area level was obtained from open-source spatial datasets. A spatial file of sub-districts was obtained from Punjab Remote Sensing Authority.

Aim and Objectives. The study's primary objectives were: (1) To describe the Spatiotemporal distribution of Dengue and its selected risk factors in Punjab, India. (2) To explore the associations between Dengue occurrence and its risk factors in Punjab, India. (3) To develop a Dengue forecasting model using routine data. We used Dengue routine data as empirical evidence for developing reproducible open-source algorithms. Thus, our secondary objective was to create a framework/algorithm for routine health data-based studies in similar settings.

Statistical methods. The study's first objective was fulfilled during the study's Exploratory Data Analysis (EDA) phase. We calculated classical summary statistical measures, dengue incidence rates, and Standardized Incidence Ratios (SIR). We also explored spatial and time series features of Dengue and its risk factors at state, district, and sub-district levels.

To accomplish the study's second objective, we undertook the data analysis and interpretation phase in a stepwise manner. We performed correlation analysis (Pearson's correlation coefficient), spatial autocorrelation analysis (*Moran's I* and Local Indicator of Spatial Association), time series analysis (Cross Correlation Function), and space-time emerging hotspot analysis (Gi* statistic and Seasonal Mann Kendall trend test). Subsequently, spatiotemporal models were developed and evaluated. Being count data, we started with Generalized Linear Models (quasi-Poisson and negative binomial models). The model complexities were increased to

enhance explainability and capture relationships on the one hand and to avoid overfitting on the other. We used Generalized Additive Models (GAMs) to capture non-linearity and Generalized Additive Mixed Models (GAMMs) to add random spatial components. The model evaluation included visualization of diagnostic plots and calculation of Adjusted R squared, Root Mean Squared Error (RMSE), and Akaike Information Criteria (AIC).

We developed and evaluated Hierarchical Time series models and GAMMs for forecasting to fulfil the third objective. Hierarchical time series forecasting models were based on Autoregressive Integrated Moving Averages (ARIMA) and time series decomposition analysis approaches. The GAMMs were trained on data from 2015-18 and tested using 2019 data. A model was considered accurate for a given location (state, district, and sub-district) when the observed number of cases during the specified period was within the 95% Confidence Interval of the forecasted value.

The additional objective to create a framework for routine health data-based studies that was considered during all the analysis phases. During the data collection, extraction, and pre-processing, we developed and used systematic, logic-based, semi-automated, and open-source reproducible algorithms for handling data quality issues to create research-level analysis-ready datasets. During subsequent phases, algorithms for spatial autocorrelation analysis, space-time emerging hotspot analysis, development of generalized models, and forecasting were developed.

Results. The total data extraction and pre-processing files for all variables included in the study were 3,858, with a data volume of 158.8 GB. For NVDBCP line-listing data, prior to data collection, we undertook coordination and exposure visits to

understand the data structure. We checked for duplicates, and the data was anonymized. A total of 66,581 case records in 67 spreadsheets were available, among which 133 were duplicates, 101 were repeat testing cases at multiple echelons of health care, and 1893 were blank records. Thus, 64,454 case records were included for further analysis.

Exploratory Data Analysis (EDA). It was observed that the annual dengue incidence in the state varied from 33.4 per lakh to 52.0 per lakh population. Further, the annual dengue incidence was significantly lower in the state during 2016 and 2019 compared to 2015 (Incidence Rate Ratio = 0.7 and 0.7, respectively). Exploration of the patterns revealed a higher dengue incidence in the last quarter of the year, the month of October, and from weeks 41-46. The mean (SD) age of the reported cases was 34.3 (16.7) years, and the maximum number of cases was in the 25-39 years age group. The majority of the reported cases (63.9%) were males. The climatic conditions suggested that 2019 was the year with the hottest and coldest days, and 2018 observed the highest precipitation. The state had high vegetation cover (average Normalized Difference Vegetation Index = 0.54) with a mean elevation of 198.8 m and a slope of 2.6 degrees.

Exploration of spatial features. Choropleth maps suggested dynamicity in annual dengue incidence across districts. The median annual dengue incidence was 36 per 100,000 population and varied from 4.8 to 210 per 100,000. The spatial distribution of risk factors such as population density, household density, urbanization, and built-up area at the sub-district level was patchy within and between districts. The north-

western region of the state had a high elevation in the Himalayan foothills, and the rest had similar elevation and low slope values.

Exploration of the time series features. The time plot of Dengue depicted the seasonal pattern of disease incidence. The dengue incidence at a given time was significantly autocorrelated with previous dengue burden at multiple lag time periods. There was a strong seasonality at monthly intervals (seasonal strength = 0.91) with a mild trend (trend strength = 0.14).

Exploration of Spatiotemporal patterns. Disease risk mapping highlighted changing epidemiology of Dengue in the state. The dengue incidence in north-western sub-districts was rising, and there was a shift to the perennial pattern in the southern sub-districts of the state. Similar to dengue incidence, the Spatio-temporal distribution of climatic and environmental factors showed seasonal patterns. There was a rising trend of temperature in the state, the southeastern sub-districts were the hottest, and the northern border sub-districts had the highest precipitation and relative humidity during the study period.

Data analysis and interpretation. Dengue incidence was correlated with multiple climatic, environmental, and socio-demographic factors. Spatial autocorrelation analysis showed clustering of Dengue in the state. *Moran's I* was statistically significant, with a positive value for multiple periods ($p<0.05$). Maximum clustering across sub-districts was present during the onset (July-Aug) and waning (Nov-Dec) of dengue season. The Local Indicator of Spatial Autocorrelation (Local *Moran's I*) showed that the location of spatial clusters was dynamic, and the sensitivity analysis highlighted the core and spread of spatial clusters.

The time series cross-correlation analysis showed a significant association of Dengue with climatic and environmental factors at multiple lags. Land surface temperature (night) was positively cross-correlated with Dengue from a lag of 2 months ($r = 0.52$, $p < 0.05$) to 5 months lag period ($r = 0.45$, $p < 0.05$). Similarly, Dengue was cross-correlated with precipitation, relative humidity, and NDVI at multiple lags. The scatter plots suggested non-linear associations of Dengue with climatic and environmental factors.

Space-time emerging hotspot analysis showed that most subdistricts were sporadic hotspots during the study period (n = 27, 18%). This was followed by the number of sub-districts that were persistent dengue hotspots during the study period (n = 21, 14%). Faridkot and Muktsar blocks in the southwestern region were persistent and intensifying hotspots. New hotspots were along the western border, namely Firozpur, Taran taran, and Gurdaspur sub-districts.

Spatiotemporal models. Based on findings from the study's data analysis and interpretation phase, we developed spatiotemporal models incorporating climatic and environmental variables at time lags with the highest correlation values. We dropped variables that were correlated with each other. Since the initial quasi-poison Generalized linear model showed overdispersion (dispersion parameter = 8.86), we shifted to a negative binomial model. The dispersion parameter of the negative binomial model was 1.05 and showed significant estimates for temperature, precipitation, relative humidity, and NDVI. However, the quantile-quantile (QQ) plot showed significant deviations from the 45-degree line suggesting non-linearity. Cook's statistics suggested high influence observation points on the model and high leverage

compared to the variance of the raw residual at a given point. The final Generalized Additive Mixed Model had reduced AIC value (dropped from 18,859 to 15,586), higher adjusted R squared (0.66), and adequate basis functions for independent variables suggesting adequate capture of non-linearity. The diagnostic deviance residual QQ plot showed uniform distribution. There were no significant patterns on the residual plot, the histogram of residuals had near-normal distribution, and there was no skewed pattern between response and fitted values. Also, there was no residual spatial and time series autocorrelation suggesting adequate capture of spatial and time series associations in the model.

Forecasting models. The bottom-up reconciliation method using the ARIMA model had the lowest RMSE values at sub-district, district, and state levels (35.9, 103.9, and 803.1, respectively) among hierarchical time series forecasting models. In comparison, the GAMM forecast model had RMSE of 23.08, 71.67, and 790.69 at sub-district, district, and state levels, respectively. The model accuracy for forecasting total dengue cases in a given month in the state was 75%, and 91.6% for hierarchical and GAMM forecasts. Further, the forecasts could identify unusual dengue occurrences across districts and sub-districts.

Strengths and limitations. Value addition to the existing routine health data using open-source, reproducible, and scalable algorithms was the major strength of the study undertaken. The major challenge in the study was related to the quality of the routine health data. We could not undertake analysis beyond the sub-district level nor point-pattern analysis to identify dengue clusters within sub-districts considering quality issues.

Conclusion and Recommendations. To conclude, the study could demonstrate the potential of using RHIS data to understand disease epidemiology and develop forecasting algorithms in resource-constrained settings. Future work should include institutionalizing data science approaches in the health systems and evaluating their potential for preventing and controlling Dengue and other infectious diseases.

,

CHAPTER 1

INTRODUCTION

1 INTRODUCTION

1.1 Research context

Dengue, an arboviral disease, is the fastest-growing mosquito-borne disease and has increased over eightfold in the past two decades globally. The risk of dengue infection is present across 129 countries, with a 70% disease burden in Asia (World Health Organization, 2022a). The World Health Assembly highlighted the importance of Dengue as an emerging disease in 2006 (Farrar and Manson, 2014). In Low- and Middle- Income Countries (LMICs), including India, Dengue is hyperendemic, a significant public health problem, and spreading to rural areas (World Health Organization, 2011). A recent nationwide serosurvey to estimate dengue infections in India found heterogeneous transmission with a high burden in the north, south, and western parts of the country (Murhekar et al., 2019). Further, according to National Vector Borne Disease Control Programme (NVBDP), the highest number of dengue cases reported in 2015 were from Punjab. More than 10,000 cases have been reported annually from the state since then (NVBDCP, 2021).

The dengue transmission dynamics include interaction between humans as a source of infection and mosquitoes as vectors for spreading the disease in the population. Multiple climatic, environmental, and socio-demographic characteristics of a community are known to affect these vector-human interactions, the vector's bionomics, and the vector's viral development process (Farrar and Manson, 2014). Among climatic factors, temperature directly affects the water source availability for mosquito breeding, survivability of development stages, mosquito reproduction rates,

and transmission and replication rates of the dengue virus (Morin et al., 2013). Water sources for breeding are further determined by the amount of rainfall, independently and closely interacting with temperature conditions. Mosquito density, an essential parameter for vector competence in dengue transmission, is influenced by micro-climatic and environmental conditions such as humidity, vegetation cover, land use patterns, and built-up area (J Xu et al., 2020; Z Xu et al., 2020; Zhang et al., 2016; Zheng et al., 2019). In favourable climatic and environmental conditions, the socio-demographic characteristics of a population further determine dengue transmission potential (Singh et al., 2019). Additionally, unplanned urbanization, inadequate public health infrastructure, lack of civic amenities, and healthcare system characteristics impact vector-human interactions (Banu et al., 2011; Ooi and Gubler, 2009).

"Systems that comprise data collected at regular intervals at public, private, and community-level health facilities and institutions and health programs" are known as Routine Health Information Systems (RHIS) (MEASURE Evaluation, 2022). These datasets provide information on the health status of the populations. The routine data is often collected for management decisions at multiple levels of health care. At the facility level, RHIS is used for micro-planning and at higher levels for strategy development, policy making, and resource allocation in LMICs (Wagenaar et al., 2016). However, despite the routine data being information-rich, it is underutilized for research, mainly due to data quality concerns. Considering the substantial efforts to strengthen the data quality of RHIS in LMICs in recent decades, the high cost of population surveys, and limited resources, there has been an increasing demand for real-time data from researchers (Hung et al., 2020).

There is no specific treatment for Dengue. Early detection and healthcare access can reduce deaths in severe cases to less than 1% (World Health Organization, 2022a). Public health surveillance aims to provide information for action at a timely interval (Last and International Epidemiological Association, 2001). The nodal programs for vector-borne disease surveillance and control in the country are the National Vector Borne Disease Control Programme (NVBDCP) and the Integrated Disease Surveillance Programme (IDSP), India (Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India, 2021). Traditionally, disease surveillance in India is an indicator-based and event-based system. Reported cases and informal information sources such as media and rumour registries form the basis for early detection and response (National Centre for Disease Control, Directorate General of Health Services, 2022). To strengthen these mechanisms, the expansion of Geographical Information Systems (GIS), linking to databases from other non-health sectors, and development of signal detection algorithms are recommended (Directorate General of Health Services, India, 2015).

Data science has emerged as a novel discipline to explore, analyze, and model ever-increasing complex routine and big-data systems (Van der Aalst, 2016). Data science is defined as "an interdisciplinary field involving processes, theories, concepts, tools, and technologies, that enable the review, analysis, and extraction of valuable knowledge and information from structured and unstructured (raw) data" (NCBI, 2021).

Increased availability of data, advancements in informatics and computational resources, enhanced use of data science applications, and resulting advances in

informatics have enabled researchers to understand the epidemiology of diseases and develop aberration detection and forecasting algorithms for disease surveillance systems (Yuan et al., 2019). Geospatial information analysis, time series analysis, and machine learning algorithms using a data science approach have been used to understand disease patterns for forecasting disease outbreaks (Bouzille et al., 2018; Chae et al., 2018; Volkova et al., 2017). Further, applying data science tools and technologies has enabled disease risk mapping in populations to efficiently utilize constrained public health resources (Minale and Alemu, 2018).

There are several research gaps in the literature related to the application of data science approaches in LMICs, especially in the Asian sub-continent, for understanding disease epidemiology, their Spatio-temporal patterns, use of routine health data, data linkages with non-health sectors and availability of the open source, reproducible and scalable algorithms. Most research using RHIS is predominantly in developed regions and, among LMICs, is limited mainly in the African sub-continent. There is a need to explore and understand the nuances of RHIS data analysis in the local context for applications in varied public health settings. The Government of India (GoI) has taken multiple initiatives to incorporate technological advancements. Integrated Health Information Portal (IHIP) was launched in a phased manner across the country; however, there is a lack of research on mechanisms by which the RHIS data can provide information and strengthen disease surveillance in India. Further, the use of excel as a data entry platform is still rampant across multiple health programs in LMICs, including India. The routine data often have spatial and temporal features; however, the lack of uniformity in dates and addresses poses limitations to understanding Spatiotemporal patterns in disease epidemiology.

Therefore, this evolution has bestowed public health researchers and epidemiologists in LMICs and India to spearhead studies on understanding mechanisms by which routine data can generate evidence for public health decision-making and strengthen disease surveillance. Thus, we undertook the present study with empirical evidence focused on routine NVBDCP dengue data from the state of Punjab. The main objectives of this research are to describe the spatiotemporal distribution of Dengue and its selected risk factors, to explore the associations between dengue occurrence and its risk factors, and to develop a dengue forecasting model using line-listing data and routine data from non-health sectors, including satellite imagery. Further, during the research process, the secondary objective is to create open-source, reproducible frameworks/algorithms for futuristic routine data-based studies.

1.2 Scope of the study

The research project adds value to the existing routine health data in LMICs. The project provides insights into mechanisms wherein routine data can be used to gather evidence for strengthening disease surveillance. The present study explores the spatiotemporal patterns of Dengue and its associations with climatic, environmental, and sociodemographic factors in the local context. Dengue forecast models developed and evaluated up to sub-district levels can help public health administrators as decision support for a timely and evidence-informed allocation of constrained resources. The development of open-source algorithms for routine health data cleaning to extract analysis-ready data with spatial and temporal features lays a foundation for future research and institutionalization of data science approach in health care systems in resource-constrained LMICs. Data linkages with routine data of non-health sectors and

satellite imagery strengthen evidence of the potential of intersectoral initiatives for precision public health.

1.3 Structure of the Thesis

The thesis outlines the research in six chapters with supporting references, annexures, and appendices. The first chapter provides an introduction highlighting the context and rationale of the research work undertaken, the scope of the study, and an overview of the structure of the thesis.

The second chapter describes the literature review as an iterative process toward establishing statistical and methodological foundations for the research. The chapter includes findings of a review on the epidemiology of Dengue, a narrative review on Routine Health Information Systems (RHIS), a bibliometric analysis that includes Citation Network Analysis (CNA) and a conceptual review of data science in infectious diseases, and a semi-automated systematic review on research undertaken for Dengue modeling/forecasting using routine data.

The third chapter of this thesis deliberates on the research methodology. This section begins by detailing the study design, geographical context of the research, study variables, data sources, and data quality characteristics. The chapter further deliberates on the methods adopted for Exploratory Data Analysis (EDA), Spatio-temporal analysis, Spatio-temporal regression models, and the development and evaluation of Dengue forecasting models.

The fourth chapter of this thesis presents the results of this research. It begins by describing the data characteristics of the routine data followed by the spatiotemporal distribution of Dengue and its associated climatic, environmental, and socio-

demographic factors up to the sub-district level in the state. Additionally, risk mapping of Dengue is presented. The chapter highlights findings of correlation analysis, spatial autocorrelation analysis, time series cross-correlation analysis and space-time emerging hotspot analysis, followed by a parsimonious development of spatiotemporal regression models. The chapter also provides findings of dengue forecasting models based on hierarchical time series and Generalized Additive Mixed Model approaches.

The fifth chapter discusses the results of this research with previous findings and emphasizes this research's importance. It also delineates the limitations and hurdles faced during this research. Another section in this chapter portrays the policy implications and essential areas of further investigation.

The final chapter provides a summary of this thesis. It provides the conclusions drawn from this research.

.

CHAPTER 2

LITERATURE REVIEW

2 LITERATURE REVIEW

2.1 Introduction

The purpose of this chapter as a literature review is as an iterative process towards establishing statistical and methodological foundations for the research. We undertook a literature review to understand Dengue epidemiology, the potential and challenges in using Routine Health Information System (RHIS) data for evidence generation, the evolution of the application of data science in infectious diseases, and existing research on dengue modeling using secondary data sources. We conducted the literature review using multiple approaches. Following a literature review on the epidemiology of dengue, we undertook a narrative review of the existing literature on Routine Health Information Systems (RHIS), a bibliometric analysis that included Citation Network Analysis (CNA) and a conceptual review of data science in infectious diseases, and a semi-automated systematic review on research undertaken for dengue modeling/forecasting using routine data. The review of the literature enabled us to understand the research gaps. It provided a detailed justification and rationale for the study undertaken.

2.2 Epidemiology of Dengue

Dengue, an arboviral disease, is transmitted by an infective female Aedes mosquito bite to humans (Farrar and Manson, 2014). The transmission of Dengue involves complex interactions between humans and mosquitoes, which are dependent on multiple factors in the host, vector, and environment. Humans are a vital source of

infection, and mosquitoes act as vectors for the transmission of Dengue in a community (World Health Organization, 2022b).

Dengue virus belongs to the genus Flavivirus and the family Flaviviridae. There are four distinct subtypes of dengue viruses (DEN-1 to DEN-4) (Farrar and Manson, 2014; Halstead, 2008). Infection from one of the subtypes confers transient immunity from infection by other subtypes. Subsequent infections from different subtypes may lead to severe forms of the disease. All four serotypes are known to cause dengue epidemics with varying severity (Farrar and Manson, 2014; Halstead, 2008; World Health Organization, 2022b). Dengue infection is often mild and asymptomatic in over 80% of cases. The clinical manifestations are often acute and flu-like illnesses; however, in some individuals, it may lead to severe Dengue, Dengue Haemorrhagic Syndrome and Dengue Shock Syndrome, and poses a substantial economic burden in Low- and Middle- Income Countries (LMICs). Further, the lack of an approved vaccines for Dengue in the majority of LMICs, including India and the lack of any specific treatment for DENV, prevention and early detection are considered as the best approach for its prevention and control (World Health Organization, 2022b).

2.2.1 Global, regional, and national burden of Dengue.

Records of illnesses resembling Dengue stretch back more than 200 years, and the dengue virus (DENV) viral aetiology was discovered in the 1940s (P M Ashburn and Charles F Craig, 2004). Dengue is the fastest-growing vector-borne disease. Globally, the number of cases has increased over eight times in the last two decades, from 5,05,430 cases in 2000 to over 5.2 million in 2019. Between 2000 and 2015, reported deaths increased from 960 to 4,032, mainly impacting younger age groups.

Although there is a risk of infection in 129 nations, the Americas, South-East Asia, and the Western Pacific are the most adversely affected regions, with Asia accounting for over 70% of the global burden (World Health Organization, 2022a).

In the South-East Asian Region (SEAR), ten dengue-endemic nations contain an estimated 1.3 billion at-risk population, contributing to about 52% of the global population. Except for North Korea, the SEAR is affected by regular and cyclical dengue epidemics. Over the past few decades, dengue virus transmission has experienced substantial growth in the SEAR, especially in India and Sri Lanka. All four Dengue strains appear to have established hyperendemic circulation, and Dengue outbreaks have increased in frequency. The countries in SEAR are divided into three categories depending upon the level of endemicity. Category A includes India, Bangladesh, Sri Lanka, Timor Leste, Indonesia, Maldives, Myanmar, and Thailand, wherein Dengue is a public health problem, hyperendemic, with all serotypes circulating in urban areas and spreading to rural areas.

In India, the dengue virus was first reported in 1944 from serum samples of U.S. soldiers in Kolkata. The first virologically proven Dengue epidemic in India occurred in Calcutta and the Eastern Coast of India in 1963-1964. The first major Dengue Epidemic in India occurred in Delhi in 1996 and subsequently spread all over the country (Gupta et al., 2012). Despite the Case Fatality Rates (CFR) decreasing from 3.3% in 1996 to 0.2% in 2018, cases are increasingly reported across the country. The Dengue virus spread throughout India by altering its genetic makeup and triggering further geographic growth. The four DENV serotypes coexist, and shifts in the dominant serotype might result in coinfection with different serotypes. DENV-1

and 2 are the most prevalent serotypes in the North region. At the same time, DENV-2 and 3 are dominant in the South region of the country. According to a nationwide multicentric sero-surveillance study, Dengue transmission in the country is hyperendemic and shows a heterogenous transmission with a high burden in the country's North, South, and West regions (Murhekar et al., 2019). The economic burden due to Dengue is the highest among vector-borne diseases in India. A recent study on decadal trends of Dengue in the country highlighted that the highest reported median annual dengue incidence is in the South, followed by the West, North, Northeast, Central, and East regions (8.18, 8.05, 4.5, 1.89, 1.62, and 1.6 per lakh, respectively). Among states, Punjab, Goa, Kerala, and Odisha had the highest median annual dengue incidence (24.49, 14.41, 12.13, and 9.1 per lakh), respectively (Singh et al., 2022).

The first reporting of Dengue cases in Punjab occurred in 1997 (23 cases and three deaths) (Gupta et al., 2012). Subsequently, the number of cases and deaths reported from the state is rising. In 2015, NVBDCP reported the highest number of cases across states and later reported more than 10,000 cases annually. The state reported less than 10,000 cases in 2020 during the COVID-19 pandemic; however, the average of 2020 and 2021 reported cases, and 2022 provisional figures followed decadal reporting trends (National Center for Vector Borne Diseases Control, 2022).

2.2.2 Association of Dengue with climatic, environmental, and socio-demographic factors.

Climatic factors are essential in dengue transmission dynamics and, thus are most often incorporated into dengue disease models. Temperature is critical in Dengue

epidemiology, affecting viral replication, vector susceptibility, vector bionomics, and vectorial competence. Increased ambient temperature leads to faster viral replication in the mosquito and a shorter External Incubation Period (EIP). A study on understanding the association between EIP and temperature in India found EIP varying from 5.6 days at 35 degree Celsius (^{0}C) to 96.5 days at zero degree Celsius in Punjab (Mutheneni et al., 2017). Diurnal Temperature Ranges (DTR) also affect the Aedes' susceptibility to DENV infections. A study estimating the impact on Dengue transmission from daily temperature fluctuations found that the susceptibility of mosquitoes to DENV infection reduces with larger DTR at the same daily mean temperatures (Lambrechts et al., 2011). Aedes being the vector of Dengue, its ecology is intrinsically tied to DENV ecology. Temperature conditions partly govern all stages of the Aedes life cycle, from egg and larval development to adult survival and mosquito feeding activities. In lab settings, Aedes development and survival rates peak at 34^{0}C and 27^{0}C, respectively (Rueda et al., 1990). The development of Aedes life stages ceases at temperatures less than 8.3^{0}C, and the ideal survival range is between 20-30 ^{0}C (Tun-Lin et al., 2000). Further, at higher temperatures, the mosquitoes become more dehydrated, and thus the rate of biting increases, resulting in increased vectorial competence.

The egg-laying, larval, and pupal stages of the *Aedes* mosquito requires clean water sources; hence, the density of the vector is also dependent on precipitation. In a study carried out in India to explore the association of rainfall with dengue, it was observed that in the northern part of the country, high rainfall was associated positively with dengue occurrence, whereas, in the southern states, the patterns were reversed. Also, the study found that the number of dengue outbreaks in rural areas was modulated by cumulative annual rainfall values (Shil, 2019). Relative humidity also

plays a vital role in the life span of mosquitoes. Adult *Aedes* survive best at a relative humidity of 60-80%. Further, the conduciveness of the environment for the transmission of dengue depends on the socio-demographic profile of the community. Poor public health infrastructure, lack of civic amenities, rapid urbanization, and other socio-demographic factors affect the availability of breeding places for mosquitoes and impact vector-human interactions (Singh et al., 2019).

2.3 Routine Health Information Systems (RHIS)

To understand the published literature on RHIS in LMICs, we conducted a narrative review on PubMed, Web of Science (WoS), Cochrane, and Google Scholar. The review provided an understanding of the RHIS, the potential of RHIS in research, data quality issues and their associated factors, and the strategies for strengthening RHIS.

A resilient healthcare system responsive to the population's needs is a prerequisite for improving the population's health. Health information, an integral part of the health system, is essential for evidence-informed decision-making, guidance to stakeholders at multiple levels, and tracking system performance (Zodpey and Negandhi, 2016). RHIS includes data collected regularly from multiple health facilities, community-level public health centres, public and private hospitals, and other healthcare institutions. It provides information on health services, health status, and resource availability in a given community (Maïga et al., 2019; MEASURE Evaluation, 2022). RHIS data is an essential tool for health system strengthening. RHIS generates multiple indicators to track national and sub-national programs toward achieving Universal Health Coverage and Sustainable Development Goals. Further,

indicators from RHIS are increasingly used to develop dashboards to enhance situational awareness and interpretation by multiple stakeholders (Maïga et al., 2019).

However, despite the routine data being information-rich, it is primarily used for administrative and managerial decisions and often overlooked for research, mainly due to data quality concerns. In recent decades, there have been substantial efforts to strengthen RHIS data quality in LMICs. Also, in resource-constrained settings, the high cost of population surveys poses a challenge to estimating health situations and program evaluations. The inherent element of repeated measurements over a long time in RHIS has led to an increasing demand for real-time data from researchers (Hung et al., 2020; Wagenaar et al., 2016).

2.3.1 Use of RHIS for research in LMICs

There has been a rising trend in using RHIS data for research purposes. In LMICs, indicators derived from RHIS lays the foundations for planning, monitoring, and evaluating health services in developing countries. Since a majority of these health indicators use population-based RHIS, it enables comparison between and within these geographical boundaries and over time (Asah et al., 2017). Additionally, the use of RHIS in combatting COVID-19 for estimating epidemiological parameters, contact tracing, policy-making, risk mapping, resource allocation, surveillance, and many other research domains highlights the potential of RHIS in evidence generation (Zhang et al., 2022). In a systematic review to understand the use of RHIS in research across 37 countries globally, more than half of the studies were published after 2014, and the majority were in the last decade. However, more than three-quarters of RHIS data-based studies were carried out in Sub-Saharan African countries. Program evaluation,

assessment of service provision, describing disease epidemiology, impact evaluations, and costing are the most common research purposes for which RHIS is used for research. Also, most studies in literature using RHIS investigate communicable diseases, especially Malaria, and ecological studies are the most common study design (Hung et al., 2020).

2.3.2 Data quality in RHIS

The data quality of RHIS has been the subject of extensive research over the years. Availability of good quality data at timely intervals is critical to data-based public health decision-making. Traditionally, based on the PRISM framework, critical factors in RHIS performance and data quality are related to environmental, organizational, technical, procedural, behavioural, and individual competence (Aqil et al., 2009; MEASURE Evaluation, 2022). In a study to understand factors associated with data quality in RHIS in Benin, it was observed that the data quality is significantly associated with responsibility level, employment sector, training received, and levels of self-efficacy and complexity perceived and supervision of a worker (Glèlè Ahanhanzo et al., 2014). In another study carried out in Odisha, India, a lack of awareness among healthcare workers regarding the importance of data being collected was attributed as a significant cause of low-quality and incomplete data (Dehury and Chatterjee, 2018). A systematic review to understand challenges in the use of RHIS data found organizational factors such as lack of adequate human and monetary support, limited laboratory support, and assumption among peripheral workers that the data is collected only for the use in ministry to be the most common issue impacting data quality (Hoxha et al., 2020). In a study carried out in Tanzania to evaluate

interventions to improve RHIS, varied methods of data collection due to multiple reporting requirements, poor internet availability, use of formats that cannot be used by other data users, and inaccuracies resulting from manual data entry were found to be related to poor data quality (Mutale et al., 2013). Further poor staff motivation, data falsification in the absence of supervisors, and lack of understanding of reports prepared centrally by the peripheral health workers have contributed to the lack of good data quality in RHIS (Asah et al., 2017; Kimaro and Twaakyondo, 2006).

2.3.3 Strengthening use of RHIS

Health Information is among one of the most important attributes of a health system which has emerged as a global and national agenda for strengthening data-driven decision-making (World Health Organization, 2007). Good quality data is paramount to the success of health information systems. Generally, data is considered high-quality if it is "fit for [its] intended uses in operations, decision making and planning while representing the real-world constructs it" (Fadahunsi et al., 2019). Harrison et al. suggest five pillars of governance, tools, processes, people, and evidence which formulate the basis for strengthening RHIS (Harrison et al., 2020). Based on the critical factors affecting data quality, measures to strengthen RHIS are organized into organizational, technical, and behavioural measures. They can be used in isolation or as a combined strategy for the improvement of RHIS and health system strengthening. In a systematic review to understand the challenges and strategies for the use of RHIS, strategies targeting organizational difficulties were found to be most effective (Hoxha et al., 2020). Infrastructural changes such as the division of RHIS responsibilities among health professionals, enhanced autonomy, and budgetary

control among peripheral health workers are known to increase data quality (Latifov and Sahay, 2013). Another study carried out in Mozambique found the inclusion of training programs for healthcare workers as an effective measure for strengthening RHIS (Wagenaar et al., 2015). Other measures, such as the development of additional cadres for supervision, leadership programs for the development of data use culture, and ranking of health facilities, have also been successful in multiple contexts in LMICs (Moyo et al., 2016; Nutley et al., 2014). Further, the majority of studies targeting behavioural factors for the improvement of RHIS have used training components and have found significant improvement in RHIS data quality at some locations (Etamesor et al., 2018; Nutley et al., 2014) and not at others (Nwankwo and Sambo, 2018).

WHO framework for RHIS for strengthening RHIS recommends including measures for all stages of data management and handling for its enhanced use in decision-making (World Health Organization, 2007). Advances in Information and Digital Technologies and data science approaches have the potential to clean and extract information from routine large datasets. RHIS data requires robust data cleaning and pre-processing before its use for research purposes. The lack of the same leads to information loss, missing data, and inaccurate outcomes when used for research (Maïga et al., 2019; Van den Broeck et al., 2005). Transparent documentation and a systematic approach are recommended; however, there is a dearth of studies that explicitly disclose the steps followed and anomalies detected and corrected during data cleaning (Maina et al., 2017; Wilhelm et al., 2019). Further, it is essential to understand data cleaning as a systematic process rather than a one-time activity. The importance of data cleaning in the data lifecycle is crucial as the resultant data's quality

would not only determine the robustness and generalizability but also allow for data linkage and sensible extrapolation of the study findings (Gesicho et al., 2020; Phan et al., 2020; Randall et al., 2013; Van den Broeck et al., 2005).

2.4 Data Science in Infectious Disease Surveillance.

In the recent past, data science has developed as a novel discipline (van der Aalst, 2016). The term data science has been defined as "an interdisciplinary field involving processes, theories, concepts, tools, and technologies, that enable the review, analysis, and extraction of valuable knowledge and information from structured and unstructured (raw) data (NCBI, 2021).

Bibliometric analysis of literature is defined as the 'application of mathematical and statistical methods to books and other media of communication'. The real breakthrough in bibliometrics occurred with the development of the Science Citation Index, i.e., a multi-disciplinary database in which authors could find articles from across many fields. The use of bibliometrics in academic medicine is in a relative state of infancy. Bibliometric parameters are playing an increasing role in academic productivity evaluation, scoring among grant applications, determining the scope for promotions among academicians, etc. Bibliometrics also aid in deciding articles, journals, and authors who are considered experts in the assessment of a given product, medicine, or equipment (Andres, 2009).

2.4.1 Citation Network Analysis

We conducted a Citation Network Analysis (CNA) using the Web of Science database to study the characteristics of available literature, understand the most

impactful sources available, and to estimate the country's scientific production globally, most cited documents, reference spectroscopy, most frequent keywords, conceptual structure, and social network of publications related to data science in infectious disease surveillance. A variety of databases exist for obtaining the information required to calculate bibliometric parameters. PubMed has become one of the most popular and widely used search engines for use with the medical literature. However, the major limitation of PubMed and Embase is that there is very little citation analysis. A large number of data fields required for bibliometric analysis are yet to be developed in the databases. As a result, choosing PubMed as well as Embase limits the application of bibliometric analysis. On the other hand, Web of Science, Scopus, and Google Scholar includes journal articles from major disciplines and have a robust citation analysis platform (Choudhri et al., 2015). The details of the search terms used are represented in appendix B.

The search identified 983 articles since 2012 with 5,108 author keywords. There were only 16 single-author documents and an average of 5.7 authors per document. The field of data science in infectious disease surveillance was found to be multi-disciplinary. It included publications from varied research categories in health care (infectious diseases, tropical medicine, public health, parasitology, immunology, and others) and non-medical branches (computer sciences, meteorology, water resources, information systems, and others). Further, based on the number of articles contributed, the top five most relevant sources were Malaria Journal, BMC Infectious Diseases, PLOS Neglected Tropical Diseases, International Journal of Environmental Research, Infection, Genetics and Evolution, and Parasites and Vectors with 141, 120, 105, 82, 65, and 60 publications respectively. The top five most cited sources were

PLOS one, Malaria Journal, Am J Trop Med Hyg, PLOS Neglected Tropical Diseases, and Lancet with 2502, 2028, 1845, 1695, and 1227 citations, respectively. The top five countries contributing to the field, based on the country of the corresponding authors, were the USA, China, Brazil, United Kingdom, and Australia, with 405, 327, 120, 78, and 74 articles, respectively.

Figure 2.1 represents the reference spectroscopy of the included articles. It highlights that the most recent work suggests the upcoming and futuristic potential of the research domain of data science in infectious disease surveillance.

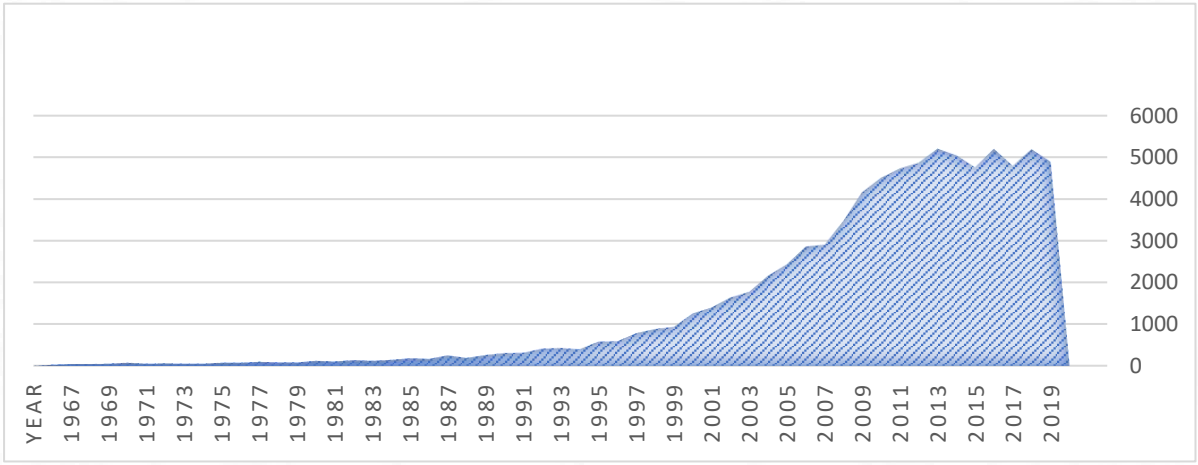


Figure 2.1 Reference Spectroscopy: Data analytics in Infectious disease surveillance.

Figure 2.2 represents the word cloud for the most frequently used author keywords. The most frequent author keywords were Malaria, Dengue, COVID-19, Influenza, spatial analysis, prediction, epidemiology, machine learning, surveillance, and pneumonia.

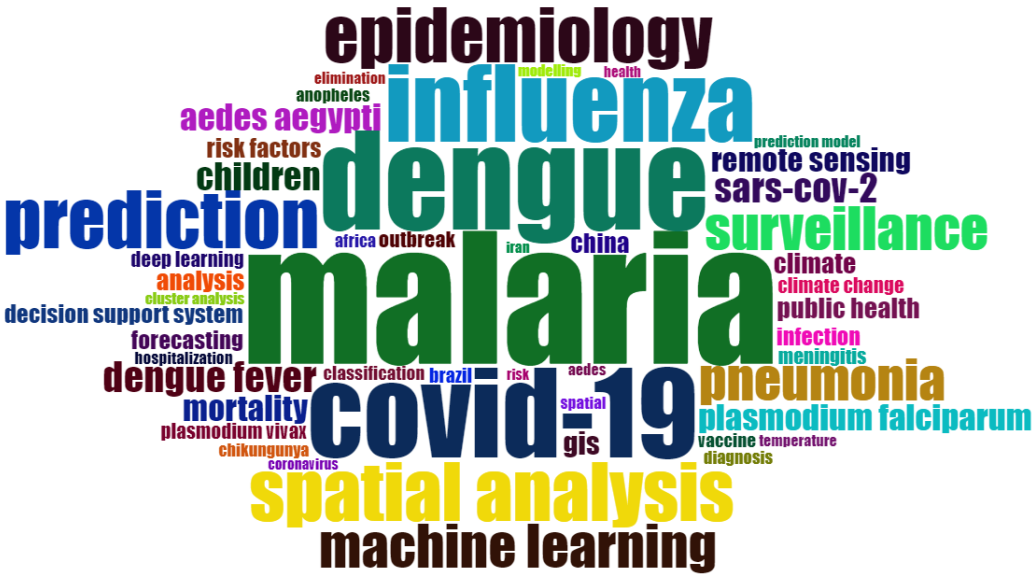


Figure 2.2 Word Cloud: Most frequent author keywords

The conceptual structure of the literature based on co-occurrence network analysis of the top 30 author keywords highlighted three clusters. The co-occurrence keywords clusters were – "remote sensing, machine learning, plasmodium vivax, malaria, GIS, climate, diagnosis, China, plasmodium falciparum, and rainfall"; "Influenza, dengue fever, epidemiology, chikungunya, forecasting, Brazil, prediction, dengue, epidemic, surveillance, transmission, Aedes aegypti, risk factors, spatial analysis, and severe dengue"; and "Pneumonia and mortality". These clusters suggest that GIS applications and spatial analysis are prominent studies carried out for understanding epidemiology of Malaria and Dengue. Also, forecasting or prediction

studies are increasingly being carried out to understand Dengue epidemics, to strengthen surveillance systems, and to explore its associations with risk factors. However, such studies are more commonly carried out in the African subcontinent.

Figure 2.3 represents the social structure of the research domain. The collaboration networks from India exist with Bangladesh, Belgium, China, Indonesia, Kenya, Nepal, Netherlands, Norway, Peru, Spain, Switzerland, and Tanzania. Similarly, collaboration networks exist from Australia, Brazil, Italy, Japan, South Africa, Pakistan, Thailand, the U.K., and the USA to India. However, as depicted in the figure, the nodes are much more prominent in the African sub-continent than in SEAR.

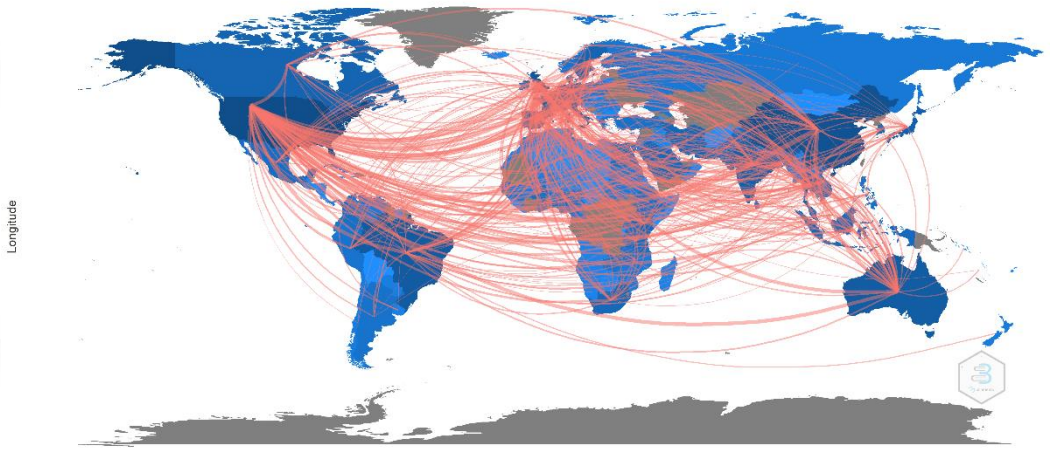


Figure 2.3 Country collaboration map: Data science in infectious disease surveillance.

2.4.2 Conceptual review

We systematically searched and reviewed the PubMed database for a conceptual review of data analytics in infectious disease surveillance. The review aimed to estimate how data science is used in infectious disease surveillance, the use patterns of data science at global, regional, national, and sub-national levels in

infectious disease surveillance, and the data sources for data science in infectious disease surveillance studies.

Data science has multiple applications in infectious disease surveillance. Forecasting disease outbreaks is a common data science application in infectious disease surveillance (Bouzille et al., 2018; Chae et al., 2018; Lowe et al., 2018; Ong et al., 2018; Pei et al., 2018; Sánchez-González et al., 2018; Volkova et al., 2017; Wang et al., 2019; Withanage et al., 2018; Zhang and Nawata, 2018). Further, data science methods have been applied to map the risk of diseases to efficiently utilize constrained public health resources (Minale and Alemu, 2018; Zambrana et al., 2018), and nowcasting using real-time data sources (Lu et al., 2019; Marques-Toledo et al., 2017). Data science methods have also been used to explore the natural history of diseases and their association with risk factors (Farinelli et al., 2018), geospatial characteristics of disease transmission (Yang et al., 2017), and to understand spatiotemporal epidemiology of diseases with implications for public health programs (Vissoci et al., 2018; Zhu et al., 2018). Data science methods have been applied to evaluate health programmes (Lopez et al., 2017) and develop location-specific operational case definitions for disease surveillance (Braga et al., 2017). Architecture for early identification of outbreaks incorporating routine health care data for public health managers has been developed using data science (Ali et al., 2016). Data science application in infectious disease surveillance has been explored in disaster scenarios, and mass screening procedures have been developed to identify persons with infectious diseases (Sun et al., 2017).

Data science is increasingly being used for infectious disease surveillance at global, national, and sub-national levels. U.S. military used data science for disease surveillance at 35 locations globally (Volkova et al., 2017). However, the majority of work in the field of data science has been applied to either national surveillance systems (Lopez et al., 2017; Lowe et al., 2018; Lu et al., 2019; Ong et al., 2018; Pei et al., 2018; Sánchez-González et al., 2018; Vissoci et al., 2018; Yang et al., 2017; Zambrana et al., 2018; Zhang and Nawata, 2018; Zhu et al., 2018) or sub-national surveillance systems (Ali et al., 2016; Bouzille et al., 2018; Braga et al., 2017; Farinelli et al., 2018; Minale and Alemu, 2018; Sun et al., 2017; Wang et al., 2019; Withanage et al., 2018). Though different datasets are being used for national and sub-national insights, web-based datasets have been used to simultaneously forecast at national and sub-national levels (Marques-Toledo et al., 2017).

Data science in infectious disease surveillance has emerged as a multi-disciplinary field in the true sense. Researchers from health background (Bouzille et al., 2018; Braga et al., 2017; Farinelli et al., 2018; Lopez et al., 2017; Lowe et al., 2018; Marques-Toledo et al., 2017; Minale and Alemu, 2018; Ong et al., 2018; Pei et al., 2018; Sánchez-González et al., 2018; Vissoci et al., 2018; Wang et al., 2019; Yang et al., 2017; Zambrana et al., 2018; Zhu et al., 2018) and non-health sectors (Ali et al., 2016; Chae et al., 2018; Lu et al., 2019; Sun et al., 2017; Volkova et al., 2017; Withanage et al., 2018; Zhang and Nawata, 2018) are contributing in the evolution of the field.

Data science in the surveillance of multiple infectious disease conditions is applied. Most notable being Dengue (Siriyasatien et al., 2018), Influenza-like illnesses

(Lu et al., 2019), Zika virus disease (Zambrana et al., 2018), Malaria (Chae et al., 2018), Measles (Yang et al., 2017), Spotted fever (Lopez et al., 2017), Viral Hepatitis (Zhu et al., 2018), and Chickenpox (Chae et al., 2018). Further, data science applications proved a powerful and essential tool in combating the COVID-19 pandemic (Zhang et al., 2022).

2.5 Research on Dengue modeling/ forecasting using routine data.

As a part of the literature review, the systematic review remains the gold standard. However, the conduct of systematic reviews remains a tedious and time-consuming process. Thus, we undertook a semi-automated systematic review using open-source algorithms as a part of the literature review. Firstly, we aim to develop a framework for the review process to reduce the manual effort and the time required for conducting an in-depth review to a minimum. Secondly, we conducted this review to estimate the global distribution of modeling/forecasting studies, identify data sources for Dengue and its risk factors, and the methods/tools used to model and forecast Dengue. The search strategy and systematic representation of the search process used for semi-automated review are presented in appendix B and C, respectively. A total of 32 articles were included in the present study to synthesize results, as represented in the subsequent paragraphs.

2.5.1 Place distribution of dengue modeling and forecasting studies.

Dengue modeling and forecasting studies are being undertaken globally; however, most studies were from Western Pacific and African regions. The highest number of included studies were from China (Liu et al., 2019; Ren et al., 2019; Zhang et al., 2019; Zheng et al., 2019; Zhu et al., 2019), followed by Indonesia (Astuti et al.,

2019; Husnayain et al., 2019; Husnina et al., 2019; Ramadona et al., 2019), Brazil (Churakov et al., 2019; MacCormack-Gelles et al., 2018; Stolerman et al., 2019), and India (Kakarla, Caminade, Mutheneni, Andrew P. Morse, et al., 2019; Swain et al., 2019; Verma et al., 2018). The probable reasons for the high number of publications from these countries, as brought forward by the authors, are the high burden of Dengue in these countries, the need for strengthening surveillance systems, and the occurrence of large-scale outbreaks. Studies from other countries included in the review were from Thailand (Phanitchat et al., 2019),(Jain et al., 2019), the Philippines (Z Xu et al., 2020), the Republic of Panama (Whiteman et al., 2019), Mexico (Sánchez-Hernández et al., 2019), Vietnam (Bett et al., 2019), Cambodia (Ledien et al., 2019), Taiwan (Anno et al., 2019), Australia (Akter et al., 2019), Bangladesh (Zahirul Islam et al., 2018),(Titus Muurlink et al., 2018), Sri Lanka (Withanage et al., 2018), Timor-Leste (Wangdi et al., 2018), Venezuela (Vincenti-Gonzalez et al., 2018), Singapore (Ong et al., 2018) and Barbados (Lowe et al., 2018).

2.5.2 Dengue data sources and pre-processing techniques.

Most studies reported dengue data obtained from national and sub-national health departments/ ministries. In addition, some researchers received data from local disease control units (Swain et al., 2019), municipality offices (Ramadona et al., 2019), data centres (Liu et al., 2019), environmental health department (Lowe et al., 2018), and weekly reports from the national surveillance programs (Kakarla, Caminade, Mutheneni, Andrew P. Morse, et al., 2019).

The studies pre-processed the reported data by either upscaling the timestamp depending on the availability of other data sources for time series analysis or undertook

data transformations to obtain linear associations and geo-referenced addresses (Ong et al., 2018) for spatial analysis. Further, depending upon the data availability, quality, and health system characteristics, either one type or a combination of patients (suspected/probable/confirmed/Dengue fever/ Dengue hemorrhagic fever) and timestamps (daily, weekly, monthly, and annual) were chosen.

2.5.3 Predictors/ Independent variables for dengue models and their data sources.

Multiple predictor/independent variables have been used in dengue modeling and forecasting. The most common variables are climatic factors: temperature, rainfall, and relative humidity. Other variables used in dengue modeling and forecasting studies included:

(a) Vector data: To directly estimate vector breeding/density/ competence leading to Dengue. However, most studies have not been able to incorporate the data on vector bionomics because of the lack of data availability.

(b) Population data including total population estimates (including population density, age distribution, sex distribution, literacy levels, piped water index, drainage index): To determine human susceptibility to Dengue.

(c) Socio-economic indicators (GDP, HDI): To determine population-level vulnerability.

(d) Normalized Difference Vegetation Index (NDVI), forest cover, land use patterns, elevation above sea level, sea surface temperature, El Niño-Southern Oscillation (ENSO) index, wind speed, road density, road networks,

distance to roads, connectivity indices: To determine ecological factors for vector breeding.

(e) Human mobility patterns: To determine the transmission chain for the spread of the disease. The data on human mobility has been used only in two studies among included studies. The inclusion of mobility data in modeling studies has been limited by data availability in the majority of study settings. Among the included studies, Ramadona et al. used geotagged tweets data as a proxy measure for human mobility in disease modeling (Ramadona et al., 2019).

(f) Internet search indices: To relate internet use behaviour with disease occurrence in a population. The use of Baidu searches in China and google trends has not been explored in LMICs owing to limited internet connectivity in the peripheral areas (Husnina et al., 2019; Liu et al., 2016).

(g) Annual homicide counts: A study carried out in Brazil for forecasting Dengue used annual homicide counts as a proxy measure for accessibility to healthcare services as an independent variable (MacCormack-Gelles et al., 2018)

2.5.4 Methods/tools used in dengue modeling and forecasting.

There are multiple approaches adopted for dengue modeling and forecasting by researchers, the most common being statistical, spatial, time-series, and machine-learning models. In the included studies, the methods used for the development of dengue models were Spatial (Ali et al., 2016; Lu et al., 2019; Minale and Alemu, 2018; Pei et al., 2018; Vissoci et al., 2018; Yang et al., 2017; Zambrana et al., 2018; Zhu et al., 2018), time series (Withanage et al., 2018; Zhu et al., 2018), Ensemble learning (Lu et al., 2019; Ong et al., 2018; Pei et al., 2018), linear regression (Bouzill et al.,

2018), multiple logistic regression (Braga et al., 2017), Generalized Additive Regression (Marques-Toledo et al., 2017), Fuzzy logic (Ali et al., 2016), mathematical (Sánchez-González et al., 2018; Wang et al., 2019), recursive and jumping prediction (Zhang and Nawata, 2018), distributed non-linear lag (Lowe et al., 2018), Principal Component Analysis (Farinelli et al., 2018), deep learning and neural networks (Chae et al., 2018; Lopez et al., 2017), discriminant analysis (Sun et al., 2017), and hierarchical cluster analysis (Yang et al., 2017) based models.

2.6 Research gaps.

While we encapsulate the literature review, many gaps seem to be unaddressed yet. Despite the fact that RHIS data strengthening is required for health system strengthening, there are numerous challenges to impede its use in practice. This has led to a distinct lack of RHIS data used for evidence generation in LMICs. Additional research is required to identify effective strategies for addressing these factors. Also, among the research projects undertaken, there is a high focus on the African sub-continent and, thus, a lack of evidence from SEAR. The present study will enable the generation of evidence on mechanisms that generate information from RHIS data using Dengue empirical data from SEAR.

Dengue is a public health problem with a high burden in SEAR. However, proactive forecasting-based models for resource allocation for its prevention and control are lacking in India. The present study will enable the development of such models for Dengue prevention using RHIS, which has the potential to generate data use culture in the health system of the country.

Dengue has been studied mostly at national and state levels in India. The present study will enable an understanding of the association of dengue with climatic, environmental, and socio-demographic factors at the sub-district level in the local context.

Further, there is a need for open-source algorithms which can handle RHIS data from LMICs. The present study is based on a data science approach and utilizes advancements in information and technology, which has the capabilities of handling unstructured and semi-structured datasets. This will enable the development of reproducible, open-source, and scalable algorithms for the present study and for use in future research.

CHAPTER 3

MATERIAL AND METHODS

3 MATERIALS AND METHODS

3.1 Introduction

This chapter provides an overview of the study design, study settings, study variables, data sources, and quality characteristics. The data science approach and database management strategies adopted to extract analysis-ready data from satellite imagery and routine health information data followed are also described in this chapter. Further, this chapter provides details of methods used for data analysis and interpretation during various phases of the study.

3.2 Objectives

The primary objectives of the study were:

1. To describe the Spatiotemporal distribution of Dengue and its selected risk factors in Punjab, India.
2. To explore the associations between Dengue occurrence and its risk factors in Punjab, India.
3. To develop a Dengue forecasting model using routine data.

We used Dengue routine data as empirical evidence for developing reproducible open-source algorithms that generate evidence from routine datasets. Thus, our secondary objective was to create a framework/algorithm for routine health data-based studies in similar settings.

3.3 Study design

The present study explored mechanisms that generate knowledge from RHIS data in resource-constrained settings. We undertook the secondary data analysis of routine health data and linked it with open data sources from the non-health sectors. The study design was an ecological study using a data science approach. We initially obtained raw data from multiple routine sources within and outside the health sector on Dengue occurrence and its associated risk factors. The raw data was pre-processed to create analysis-ready tidy data. These pre-processed tidy datasets were used to understand Dengue epidemiology. Development and evaluation of statistical and forecasting models were carried out to generate empirical evidence.

3.4 Data science approach and data analysis plan

The data science approach used included the following phases of the study: -

(a) Database management for data collection, extraction, preprocessing, and data linkages.

(b) Exploratory Data Analysis (Objective 1 and 2).

(c) Data Analysis and Interpretation (Objective 2).

(d) Development and evaluation of Dengue forecasting models (Objective 3).

The additional objective was to create a framework for routine health data-based studies that were considered during all the analysis phases. The framework and algorithms are represented in appendix E of this thesis.

3.5 Study setting

The study was carried out in Punjab state, which is in the northern part of India. Punjab has an area of 50,362 square kilometres (sq. km) with an average population density of 551 persons per sq. km. The state has 22 districts and 150 blocks. The two new districts formed since Census 2011 include *Fazilka* and *Pathankot*, which were formed on 27 July 2011 by the partition of *Firozpur* and *Gurdaspur* districts, respectively. **Figure 3.1** represents the study settings as an inset map.

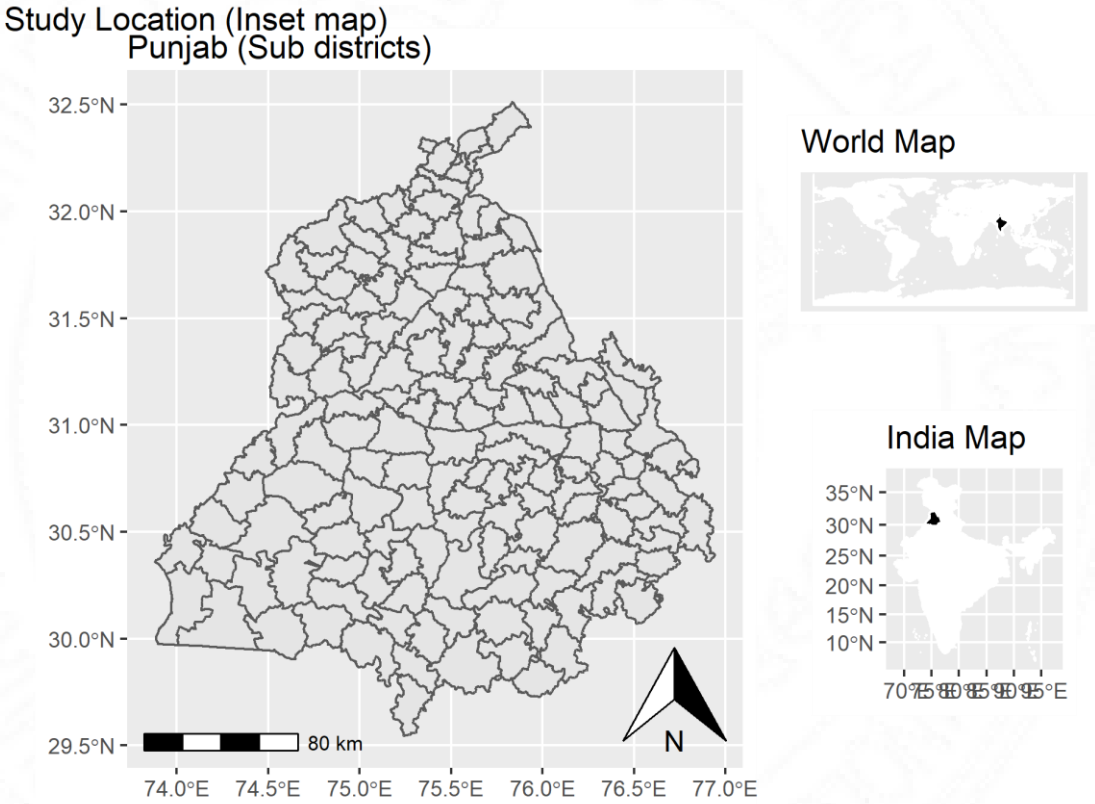


Figure 3.1 Inset map of sub-districts of Punjab, India

3.6 Study variables, data sources, and their characteristics

3.6.1 Dengue epidemiological data.

The data source for Dengue epidemiological data was line listing data maintained by the National Vector Borne Disease Control Programme (NVBDCP) at the Directorate of Health Services, Punjab. A confirmed case of Dengue is "A case compatible with the clinical description that is laboratory confirmed." Laboratory criteria for the diagnosis of Dengue include one or more of the following:

(a) Isolation of the Dengue virus from serum, plasma, leucocytes, or autopsy samples

(b) Demonstration of a fourfold or greater change in reciprocal IgG or IgM antibody titers to one or more Dengue virus antigens in paired serum samples.

(c) Demonstration of Dengue virus antigen in autopsy tissue by immunohistochemistry or immunofluorescence or in serum samples by ELISA.

(d) Detection of viral genomic sequences in autopsy tissue, serum, or CSF samples by polymerase chain reaction (PCR).

As applicable, the study variables obtained from dengue line listing datasets included registration date, testing date, and admission and discharge dates. Also, the line listing data included details on the patient's residential address, age, and gender. The datasets were in excel format with semi-structured formats.

3.6.2 Climatic data.

Temperature. The study variables on climatic data were obtained from various satellite imagery and global climatic modeling datasets. The temperature

variables included in the study were diurnal Land Surface Temperature (LST) data (day and night) from Moderate Resolution Imaging Spectroradiometer (MODIS) Land Surface Temperature and Emissivity (LST&E) data Version 6 (MOD11C1 ver 6.0). MOD11C1 ver 6.0 datasets were research-level datasets providing temperature values in Kelvin. The cell values had a valid range of 7500-65535 at a spatial resolution of 0.05 degrees (approx. 5 km) and temporal resolution of eight days. Additionally, a scale factor of 0.02 was used for the creation of MOD11C1 Hierarchical Data Format (HDF) files for efficient storage.

Precipitation. Precipitation data was obtained from the research-level Integrated Multi-satellitE Retrievals of the Global Precipitation Mission (IMERG Final Precipitation L3). The precipitation data was provided in millimetres at a spatial resolution of 0.1 degrees (approx. 10 km) daily. A scale factor of 10 was used for the creation of IMERG NetCDF (network Common Data Form) files.

Humidity and additional variables. Data on Relative Humidity, Specific Humidity, minimum and maximum temperatures at 2 and 10 meters, dew point temperature, and Wet bulb temperature were obtained from the Prediction of Worldwide Energy Resources (POWER) Project. The datasets from the POWER project were available at a spatial resolution of 0.5 degrees (approx. 50 km) daily. These datasets were based upon Goddard's Global Modeling and Assimilation Office (GMAO) Modern Era Retrospective-Analysis for Research and Applications (MERRA-2) climatic assimilation model.

3.6.3 Environmental data

Vegetation. The data on vegetation cover measured as Normalized Difference Vegetation Index (NDVI) was obtained as a filtered NDVI Product provided by National Remote Sensing Centre from Ocean Color Monitor Version 2 (OCM2) satellite imagery data. The data was available at a spatial resolution of 1080 meters for fortnightly intervals. The data was available in GeoTIFF file format with a valid range of 0-200 and an image background value of 255.

Elevation and slope. Data on elevation and slope was obtained from the national digital elevation model created from Cartosat-1 satellite imagery data (CartoDEM ver 2.0) provided by National Remote Sensing Centre, India. The data was available as GeoTIFF files at a spatial resolution of 30 meters.

3.6.4 Socio-demographic data

Socio-economic variables. Population density, household density, female literacy rates, and persons per household were obtained from the National census survey conducted by the Census Organization of India (Census 2011).

Urbanization and built-up area. Urbanization and the built-up area were obtained from high-resolution urban classification data for India for 2011. The raw spatial data contained gridded estimates at 1 km resolution with two spatial renderings based on Census 2011 tabulation for settlement types and remotely sensed built-up measures from Global Human Settlement Layer (GHSL 2014).

Projected populations. The projected population of a given sub-district was calculated using Census 2001 data, Census 2011 data, India Census, v1 (2011) spatial

datasets. The spatial dataset contained gridded estimates at 1 km resolution. Subsequently, the calculated projected populations were adjusted to state projections provided by the National Commission on Population.

3.6.5 Spatial boundaries.

The spatial file containing sub-district and district boundaries was obtained from Punjab Remote Sensing Authority. The spatial file was provided in shapefile format.

3.7 Database management strategies

3.7.1 Data collection strategies.

3.7.1.1 NVBDCP data.

Understanding data management in the field. We undertook coordination visits and short-term exposure visits to the National Centre for Disease Control, New Delhi; Directorate of Health Services, Punjab; National Vector Borne Disease Control Programme, India; Integrated Disease Surveillance Programme, India and the state health departments to understand the process of data collection and storage. Dengue is a notifiable disease in Punjab. All the lab-confirmed cases are reported to the Directorate of Health Services, Punjab.

Pre-data collection activities. We checked for duplication and repeated test cases based on name, age, sex, contact number, address, year, and date of testing before anonymization of the datasets. Line listing data of these lab-confirmed dengue patients from 01 January 2015 to 31 December 2019 were then anonymized using '*epitrix'* package in R software and used subsequently for the present study.

3.7.1.2 Climatic data.

Understanding data management in the field. We undertook visits to State Meteorological Department and processed web-platform-based requests available from Indian Meteorological Department (IMD) (Indian Meteorological Department, 2022) at http://dsp.imdpune.gov.in/ to collect climatic data from weather stations in the state. We also explored IMD gridded data for rainfall and temperature. However, considering the low spatial and temporal resolution of the on-ground meteorological stations in the state (Three meteorological stations at Amritsar, Ludhiana, and Patiala with patchy information due to equipment dysfunction or other reasons), the low spatial resolution of the gridded data (temperature data at 0.25*0.25 degree grid and rainfall data at 1*1 degree grid which is approx. 25 and 100 km^2 grids respectively), and Punjab being a state spanning from foothills to near desert areas, the discussion with experts recommended that the interpolation techniques shall not be appropriate with basic models, and thus climatic and environmental variables be collected from multiple satellite imagery sources and climatic models.

Data collection strategy for temperature and precipitation data. Since both MODIS and IMERG data repositories are available from Distributed Active Archive Centers (DAACs) of NASA's Earth Observing System Data and Information System (EOSDIS), user registration and profile management was carried out at https://search.earthdata.nasa.gov/. Subsequently, we used mass downloading scripts for data download through the Unix-like command line utility “Cygwin” for API-based extraction.

Data collection strategy for humidity and other climatic variables. We adopted Application Programming Interface (API) algorithms provided by '*nasapower'* package in R software.

3.7.1.3 Environmental data

Environmental data on NDVI, elevation and slope was downloaded from the Open data archive of Bhuvan, Indian Geo-platform, under the theme "land-Vegetation" and "land and terrain" respectively. The source website used was https://bhuvan-app3.nrsc.gov.in/data/download/index.php#.

3.7.1.4 Socio-demographic data

The census data was downloaded as Census 2011 tables from https://censusindia.gov.in/census.website/data/census-tables. The India-spatial files containing Census 2011 population and urbanization data were downloaded from Open source datasets provided by Socioeconomic Data and Applications Center (SEDAC) hosted at Columbia University website https://sedac.ciesin.columbia.edu/data/set/india-spatial-india-census-2011/data-download.

3.7.2 Data extraction and pre-processing

3.7.2.1 NVBDCP data.

We followed the framework provided by Broeck et al. to extract and pre-process the NVBDCP datasets. The framework recommends data cleaning as a process with components of screening, diagnosis, and editing (Van den Broeck et al., 2005). *A priori* operational definitions, semi-automated reproducible algorithms, and

computational workflows were created. The prepared algorithm systematically screened NVBDCP datasets, and an automated algorithm cleaned all the data anomalies fitting *a priori* definitions. A manual correction was carried out for the strange patterns with valid implicit values. In case of failure to obtain any valid value, the respective data cell was labelled as missing data. Missing data imputation based on relational variables such as dates of testing, admission, and discharge was also carried out. The detailed algorithm and workflows were peer-reviewed and published in the Proceedings of the 1st Virtual Conference on Implications of Information and Digital Technologies for Development, 2021 and are represented in appendix D.

3.7.2.2 Satellite imagery datasets.

We developed and used open-source reproducible, scalable, and automated algorithms from satellite imagery data for spatiotemporal climatic risk assessment. The algorithms need to be provided inputs regarding details of the location/directory/path of the stored files and the multi-polygon file (Punjab spatial file in the present study), following which it automatically reads a satellite imagery file in complex formats (HDF/NetCDF), extracts longitude and latitude details, extract data, applies scale factor and offset values for conversions if required, creates a raster and stores the raster file in a sub-folder for further use. Additionally, the algorithms load the multi-polygon file, transform its CRS to the CRS of the raster, crop the raster up to one pixel outside the polygon extent, and perform zonal statistics to extract polygon-wise climatic data. The process is automatically then iterated over all the satellite imagery files and provides an analysis-ready excel output for further analysis. The generic algorithms

for the MODIS, IMERG, CartoDEM, and API-based NASAPOWER data extraction are provided in appendix E.

3.7.2.3 Urbanization and built-up area

Level of urbanization was defined as a percentage area within the subdistrict which is classified as Urban according to Census 2011. The percentage built-up area for a subdistrict was defined based on GHSL-based classification of an area for both 1% and 50% thresholds as percentage area within the subdistrict which had built-up area above the specified threshold area.

3.7.2.4 Population projections

To calculate population projections, we used the formula for obtaining exponential growth rates as under:

$$P_t = P_0 * e^{rt}$$

wherein P_t is the present population; P_0 is the base year population; r is the annual growth rate, and t is the duration between baseline and the year under consideration. To calculate the growth rate, the following formula was used: -

$$r = (Pop_{2011i} - Pop_{2001i})/10$$

wherein Pop_{2011i} is the population of a given district *i* as per Census 2011 and Pop_{2001i} is the population as per Census 2001. The calculated growth rates for a given district were applied to the district and its sub-districts for calculating the projected population. The population of the sub-districts was extracted from the India Spatial file based on Census 2011 using zonal statistics. The calculated projected population

of the districts was then adjusted to the state level Projected populations provided by the National Commission on Population

3.7.3 Strategies for Data linkages

Multiple data linkage strategies were adopted for database management and for the creation of a master dataset for analysis. The data were transformed into long and wide formats for data linkages as required. Unique identifiers were identified, such as district names for attribute-based relational joins, and spatial joins were used for geometry-based operations. In spatial data handling, deliberations on the same Coordinate Reference System were carried out and ensured before linking datasets. In time series analysis, upscaling/downscaling of datasets was carried out to generate data with the same timestamps. In spatiotemporal analysis, space-time cube models were created, which included a spatial slice at each timeline and a time bin for each spatial unit.

3.8 Exploratory Data Analysis (EDA)

Exploratory Data Analysis was carried out to estimate the Spatiotemporal distribution of Dengue and its climatic, environmental, and socio-demographic risk factors at the state, district, and sub-district levels. Dengue incidence rates were calculated using the formula: -

$$Inci_{ij} = \left(Cases_{ij}/Population_{ij}\right) * 100{,}000$$

wherein $Inci_{ij}$ is the dengue incidence per 100,000 population for a given area i (state, districts and sub-districts, in a sequential manner) during the period j, $Cases_{ij}$ is the

number of lab-confirmed dengue cases reported from that geographical region during the period under consideration and $Population_{ij}$ is the projected population.

We calculated annual, quarterly, monthly, and weekly dengue incidence rates up to the sub-district level for further analysis. Descriptive statistics as percentages and mean (SD) were calculated, and the age variable was categorized into age groups to describe the characteristics of reported cases. For visualization of the spatial and spatiotemporal distribution of dengue incidence, choropleth maps and hovemoller diagrams were plotted. Similarly, to explore the spatiotemporal distribution of climatic, environmental, and sociodemographic factors, descriptive statistics were calculated, and time plots and choropleth maps were created.

To understand the time series features of dengue, auto-correlation, and partial-autocorrelation coefficients were calculated and visualized. Also, time series data of dengue occurrence was decomposed into seasonal, trend, and remainder components. Hurst coefficient as a measure of 'long memory' and spectral entropy as a measure of 'noise' were calculated for annual, quarterly, monthly, and weekly time series dengue data. Crude Standardized Incidence Ratios (SIR) were calculated to estimate the distribution of dengue risk in space and time across sub-districts using the following formula: -

$$SIR_{it} = O_{it}/E_{it} = O_{it}/\lambda P_{it}$$

wherein, SIR_{it} = Standardized Incidence Ratio for i^{th} sub-district at time t, O_{it} = Observed number of cases for i^{th} sub-district at time t, E_{it} = Expected number of cases for i^{th} sub-district at time t, λ = Relative risk of Dengue in the state, and P_{it} = population of i^{th} sub-district at time t

Sub-districts with SIR more than one were considered at high risk of Dengue in the state during the time under consideration.

3.9 Data Analysis and Interpretation

We carried out correlation analysis, spatial autocorrelation analysis, time series cross-correlational analysis, and space-time emerging hotspot analysis. Correlation analysis included the calculation of Pearson's correlation coefficients between various independent variables to assess collinearity before the development of regression-based models. Correlational plots and matrices were created for visualization of the results. Time series cross-correlational analysis included calculation of cross-correlation coefficients to estimate the lag associations between dengue occurrence and climatic and environmental variables in time series

3.9.1 Spatial autocorrelation analysis

Spatial autocorrelation analysis was carried out based on the neighbourhoods matrix which was created based on the queen's contiguity. Polygons were considered neighbours when they shared a boundary line or a point. Row standardized weight matrix was created. We calculated *Moran's I* statistic as an estimate of Global clustering using the following formula: -

$$I = \frac{N\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij})\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

wherein N is the total number of dengue cases, X_i is the number of dengue cases at a given location, X_j is the number of dengue cases at another location, W_{ij} is the weight

provided for the another location and $\overline{X}$ is the mean number of dengue cases in the state.

A *Moran's I* value of 0 was considered as Complete Spatial Randomness. Any value more than 0 was suggestive of Spatial clustering, and less than 0 suggested dispersion. Local estimates of spatial clustering were based on Local *Moran's I*, which categorized each sub-district among low-low, high-high, high-low, and low-high quadrants. A sub-district was considered as low-low when the dengue incidence was significantly low in the district under consideration, and significantly low incidence was observed in the neighbouring sub-districts as compared to the state incidence rate. A high-high subdistrict had significantly high dengue incidence in the sub-district as well as in neighbours. A sub-district was categorized as high-low when the dengue incidence was significantly high in the given sub-district but had a significantly low incidence in the neighbours, and the reverse was true for low-high sub-districts. The reproducible algorithm for spatial autocorrelation analysis is represented in appendix E.

3.9.2 Space-time emerging hotspot analysis.

We created a space-time cube for emerging hotspot analysis. Gi* statistic and Seasonal Mann-Kendall tests were performed to classify the subdistricts based on definitions adapted from ArcGIS literature. The operational definitions used in the present study are elaborated in

Table *3.1* and the reproducible algorithm is represented in appendix E.

Table 3.1 Operational definitions: Space-time emerging hotspot analysis.

S.No.	Type	Operational definitions
1	Persistent and intensifying	A location which is a statistically significant hotspot for at least four years during the study period. Additionally, the intensity of clustering has a statistically significant increasing trend.
2	Persistent	A location which is a statistically significant hotspot for at least four years during the study period. However, the intensity of clustering does not have a statistically significant trend.
3	Persistent and diminishing	A location which is a statistically significant hotspot for at least four years during the study period. Additionally, the intensity of clustering has a statistically significant decreasing trend
4	Emerging	A location which is a statistically significant hotspot for at least two years during the period 2017-2019 but has never been a statistically significant hotspot earlier.
5	New	A location which is a statistically significant hotspot in 2019 but had never been a hotspot earlier
6	Oscillating	A location which has remained a statistically significant hot spot in an on-off fashion during the study period
7	Historical	Not a hotspot in the last two years but a hotspot for at least two years during the previous three-year period
8	Sporadic	A location which has remained a hotspot for at least one year during the study period.
9	Not categorised	A location which does not meet any criteria mentioned above. Includes those areas which have never been statistically significant hotspots during the study period.

3.9.3 Spatiotemporal models

3.9.3.1 Generalized Linear Models (GLMs)

The generalized spatiotemporal regression model development was carried out in an iterative and parsimonious manner. Being a count dataset, Poisson regression and its extensions were considered. Based on the exploratory data findings, non-linear

patterns were revealed; thus, generalized linear model approach was adopted at the outset, assuming a Poisson distribution of Dengue. However, due to the high overdispersion statistic obtained, a negative binomial distribution-based modeling which is an extension of Poisson generalized linear models adjusting for overdispersion, was adopted for further analysis. Spatial distribution was incorporated at the sub-district level as it was the most granular geographical unit available for analysis. Temporally, monthly aggregated values were used to build spatiotemporal models. Additionally, an offset term for projected population values were included to adjust for the varying populations across space and time in the subdistricts during the study period.

The Quasi-Poisson regression model was based on the formula: -

$$count_i \sim \text{quasipoisson}(\mu_i, \phi)$$

$$Y_{it} = \beta_0 + \sum \beta_{it} x_{i.lag_n} + \sum \beta_j z_i + log(popualtion_i)$$

wherein $x_{i.lag_n}$ were climatic variables and NDVI at specified lags and z_i were environmental and sociodemographic factors. Overdispersion was estimated using the following formula: -

If E(Y) = μ, the quasi-Poisson model assumes var(Y) = θμ where θ is the dispersion parameter calculated as

$$\theta = \sum (Pearson\ residuals) 2/n-p$$

wherein n is the number of observations and p is the number of parameters.

The negative binomial model was based on the following formula: -

$$count_i \sim NegBin(\mu_i, \phi)$$

$$Y_{it} = \sum \beta_j x_{jit} + \sum \gamma_j w_{ji} + log(population_i)$$

wherein Y_{it} is the dengue count for a given location *i* at time *t* , $\sum \beta_j x_{jit}$ is for climatic variables and NDVI for location *i* at time *t* and j are specified lags, and $\sum \gamma_j w_{ji}$ is the for environmental and sociodemographic factors. The reproducible algorithm for GLMs is represented in appendix E.

3.9.3.2 Generalized Additive models

Generalized additive models (GAMs), a non-parametric approach that is an extension of GLMs, were adopted for further analysis. This was based on findings of GLMs which seemed inadequate to capture the complex nature of Dengue incidence and its interaction with risk factors. A baseline generalized additive spatiotemporal model was prepared by incorporating sub-districts as spatial components with monthly aggregated values for the variables to represent temporality.

Temperature, rainfall, and humidity at a lag of 1-3 months were assessed iteratively to select the optimal lag values for model building. Additionally, wind speed and NDVI at one month lag and socio-demographic variables were included in the model. The model iterations were based on the following formula: -

$$count_i \sim NegBin(\mu_i, \phi)$$

$$Y_{it} = \sum s(\beta_j x_{jit}) + \sum s(\gamma_j w_{ji}) + log(population_i)$$

wherein Y_{it} is the dengue count for a given location *i* at time *t* , $\sum s(\beta_j x_{jit})$ is the smooth function for climatic variables and NDVI for location *i* at time *t* and j is the specified lag, and $\sum s(\gamma_j w_{ji})$ is the smooth function for environmental and sociodemographic factors. The optimal lag chosen was based on the lowest AIC values. Basis dimensions check using k index was carried out to determine the optimal dimensionality of the spline basis functions for a given model. Smooth plots depicting the wiggly distribution of the variable values on x-axis were made to explore the potential improvements for the models. For the variables with gaps on x axis, transformation into categorical variables was carried out after calculating 33rd and 66th percentiles for the distribution of the respective variable. To this end, elevation was categorized as low, medium, and high; urbanization as predominantly rural, semi-urban, and urban; and household density and female literacy rate as low, medium, and high. Further, to capture seasonality, the month was added as a seasonal factor. Based on multiple iterations for the k basis functions, categorization of non-climatic factors for better spatiotemporal models, a model with a two-month lag for climatic factors with increased basis functions and socio-demographic factors as a categorical variable was developed. Further, to incorporate the effects of both minimum and maximum temperature, a temperature range was added. Random effects were incorporated for sub-districts. Repeated model diagnostics were performed to select the best-fit model. The formula for GAMM was as under: -

$$count_i \sim NegBin(\mu_i, \phi)$$

$$Y_{it} = \sum s(\beta_j x_{jit}) + \sum s(\gamma_j w_{ji}) + log(population_i) + re(sub\text{-}district) + cyclic(months)$$

wherein Y_{it} is the dengue count for a given location *i* at time *t* , $\sum s(\beta_j x_{jit})$ is the smooth function for climatic variables and NDVI for location *i* at time *t;* j is the specified lag; $\sum s(\gamma_j w_{ji})$ is the smooth function for environmental and sociodemographic factors, and *re(sub-district)* is the random component for a given subdistrict. A total of 1000 simulations were carried out to estimate the coefficients. The reproducible algorithm for GAMMs is represented in appendix E.

3.10 Dengue forecasting models

We used multiple approaches for the development of dengue forecasting models. Hierarchical time series forecasting approaches were used on the one hand, and the GAMM-based approach by dividing the dataset into training and testing data on the other.

The development of a hierarchical time series forecasting model was based on both Autoregressive Integrated Moving Average (ARIMA), and Time-series decomposition (ETS) approaches being complementary to each other. The generic computational workflow included the creation of a monthly time series dataset of block-level Dengue occurrence followed by hierarchical level definition/ aggregation at block and district level as under: -

$$Y_{State_t} = \sum y_{district_{it}}$$

$$Y_{district_{it}} = \sum y_{sub-district_{jit}}$$

Subsequently, the respective hierarchical model was computed using the hierarchical dataset, and base forecasts were reconciled to obtain coherent one-month point forecasts. The base forecasts generated were reconciled using bottom-up, top-down, middle-out, ordinary least square regression, and trace minimization methods

for comparison. Root Mean Square Error (RMSE) values were calculated to measure model performance for forecasting using the following formula to select the best-performing model.: -

$$RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

wherein, y_i = Reported cases in the selected administrative area for the i^{th} month; $\hat{y}_i$ = forecasted cases in the selected administrative area for the i^{th} month; and N = number of administrative areas under consideration. . Further, 95% confidence intervals were calculated for the forecasted values, and visualizations for each district and sub-district were prepared and analysed. The reproducible algorithm for hierarchical time series forecasting is represented in appendix E.

GAMM approach for forecasting used data from 2015 to 2018 as training data with iterations on the lag period. The best-fit model was chosen based on Akaike information criterion (AIC) and tested using the dataset of 2019. The model diagnostics included an assessment of accuracy defined as the occurrence of the number of cases in a given location for the specified month within 95% of the forecast value, AIC, and the creation of diagnostic plots. Residual spatial and time series autocorrelation analysis was performed to assess the adequacy of capturing spatial and time series components into the model. The reproducible algorithm for GAMM-based forecasting is represented in appendix E.

3.11 Statistical and Machine learning software

All the analysis was carried out using R version 4.0.3 (R Core Team, 2020) and above.

3.12 Ethical considerations

The study was carried out after obtaining Institutional Ethics Committee clearance (IEC/IEC-1653; IEC Reg No. ECR/189/Inst/KL/2013/RR-16). The research has been registered under the Clinical Trials Registry of India (CTRI/2021/01/030245). We obtained permissions for NVBDCP data use from the Directorate of Health Services, Punjab. Sub-district level spatial boundaries file was obtained from Punjab Remote Sensing Authority, and permission for its use for this research project was obtained.

CHAPTER 4

RESULTS

4 RESULTS

This chapter presents the research findings in the following manner: -

1. Data characteristics of routine datasets.
    a. Raw data characteristics
    b. Data quality features.
2. Exploratory Data Analysis (Objective 1 and 2).
    a. Exploration of Dengue dynamics in the state.
    b. Exploration of Climatic and Environmental factors.
    c. Spatial distribution of Dengue.
    d. Spatial distribution of socio-demographic factors.
    e. Spatial distribution of environmental factors.
    f. Time series features of Dengue.
    g. Spatio-temporal epidemiology of Dengue.
    h. Spatio-temporal disease risk mapping of Dengue.
    i. Spatio-temporal distribution of climatic and environmental factors.
3. Data analysis and interpretation (Objective 2).
    a. Correlation analysis.
    b. Spatial auto-correlation analysis.
    c. Time series cross-correlation analysis.
    d. Space-time emerging hotspot analysis.
4. Spatiotemporal models (Objective 2 and 3).
    a. Generalized Linear Models (GLMs).
    b. Generalized Additive Models (GAMs).

c. Generalized Additive Mixed Models (GAMMs).

5. Dengue forecasting (Objective 3).
    a. Forecast based on Hierarchical Time Series Model.
    b. Forecast based on Generalized Additive Mixed Model.

4.1 Data characteristics of routine datasets

NVBDCP line-listing raw data was available as 67 separate excel sheets. The line listing data was collated district-wise for 2015, 2016, and 2018 whereas data from all districts were available in a single state-level excel sheet for 2017 and 2019. The raw data included 66,581 rows/ observations with 1893, 133, and 101 blank, duplicates, and repeat testing records respectively.

The unique anonymized dataset included 64,454 rows. The date of registration/ admission, testing, and discharge was available in 32,472 (50.4%), 57,505 (89.2), and 9,727 (15.1%), respectively. The duration of hospitalization was available for 7,878 (12.2%) records. The columns for age 'AND'/'OR' gender details were filled for 63,717 (98.8%) records but were untidy in 9,832 (15.3%).

The logic algorithm for data extraction screened a total of 2,04,985 cells to contain date values. The excel-numeric format was present in 42,902 cell values. Among 1,31,931 values identified as editable dates, the *a priori* cleaning codes (automated) processed 1,31,569 (99.7%) cells, and 362 (0.3%) values required manual correction. The algorithm estimated testing date for 96.1% and 98.9% of observations for 2015 and 2016, respectively, and for all cases in 2017, 2018, and 2019. Age extraction varied from 99.4% (2017 and 2019) to 98.4% (2015) from available records annually. The algorithm for gender was successful in 63,006 (97.7%) cases, and

location details were available for all the reported cases during the study period. The geocoding of standardized addresses was successful in 63,741 (98.8%) records. Therefore, after pre-processing NVBDCP data, the missing percentage for 2015-19 was 1.8, 0.8, 0.3, 1.4, and 1.3%, respectively.

The total data files pre-processed to create analysis-ready data included 3858 files with a data volume of 158.8 GB. **Table 4.1** represents the data characteristics for the pre-processed files for this research.

Table 4.1 Data Characteristics

S. No.	Variable	Source	Number of files for processing	Raw data Volume
1	Dengue occurrence	NVBDCP, Punjab	67	20.3 MB
2	Temperature	MOD11C1 Version 6 (LST&E)	1,826	74.3 GB
3	Precipitation	GPM IMERG Final Precipitation L3	1,826	83.6 GB
4	Multiple climatic data*	MERRA-2	API Based extraction	
5	Vegetation	OCM2: Filter NDVI Product	118	326 MB
6	Urbanization	SEDAC	1	470 MB
7	Built-Up Area	SEDAC		
8	Elevation	CartoDEM ver 2.0	19	124 MB
S. No.	**Variable**	**Source**	**Number of files for processing**	**Raw data Volume**

9	Slope	CartoDEM ver 2.0		
10	Projected Population	Calculated	-	-
11	Spatial Datasets	PRSC, Ludhiana	1	1.31 MB
	Total		**3,858**	**158.8 GB**

* *Air temperature, Relative humidity, Specific humidity, Wet-bulb temperature, Total precipitable water vapour, Dew point temperature at 2-meter and 10-meter

4.2 Exploratory Data Analysis (EDA)

4.2.1 Exploration of Dengue dynamics in the state.

Table 4.2 illustrates the annual dengue incidence rates in the state. The yearly median dengue incidence during the study period was 47.8 cases per lakh population. The incidence rates in 2017 and 2018 were significantly higher compared to 2015.

Table 4.2 Annual Dengue incidence rates in the state

Year	Cases	Projected population	Incidence (per lakh)	IRR[1]	95% CI[2]
2015	13,829	28,954,000	47.76	Ref	-
2016	9,831	29,220,000	33.64	0.71	0.69, 0.73
2017	15,329	29,460,000	52.03	1.11	1.08, 1.13
2018	14,763	29,699,000	49.71	1.07	1.04, 1.09
2019	9,989	29,939,000	33.36	0.72	0.70, 0.74

[1]IRR = Incidence Rate Ratio, CI = Confidence Interval

Figure 4.1 represents cumulative and annual dengue occurrence during the state's quarterly, monthly, and weekly timestamps**.** Quarterly dengue occurrence ranged from 0 to 13193 cases, with a median of 571 and a maximum of patients reported in the fourth quarter. Monthly dengue occurrence ranged from 0 to 7336, with a median of 30 patients. The peak month of dengue occurrence during the study period was October, followed by November and September. On a weekly timestamp, the

highest cumulative cases were reported in week 45, and the peak reporting week ranged between week 41 and week 46.

Figure 4.2 represents the age distribution of dengue cases reported across

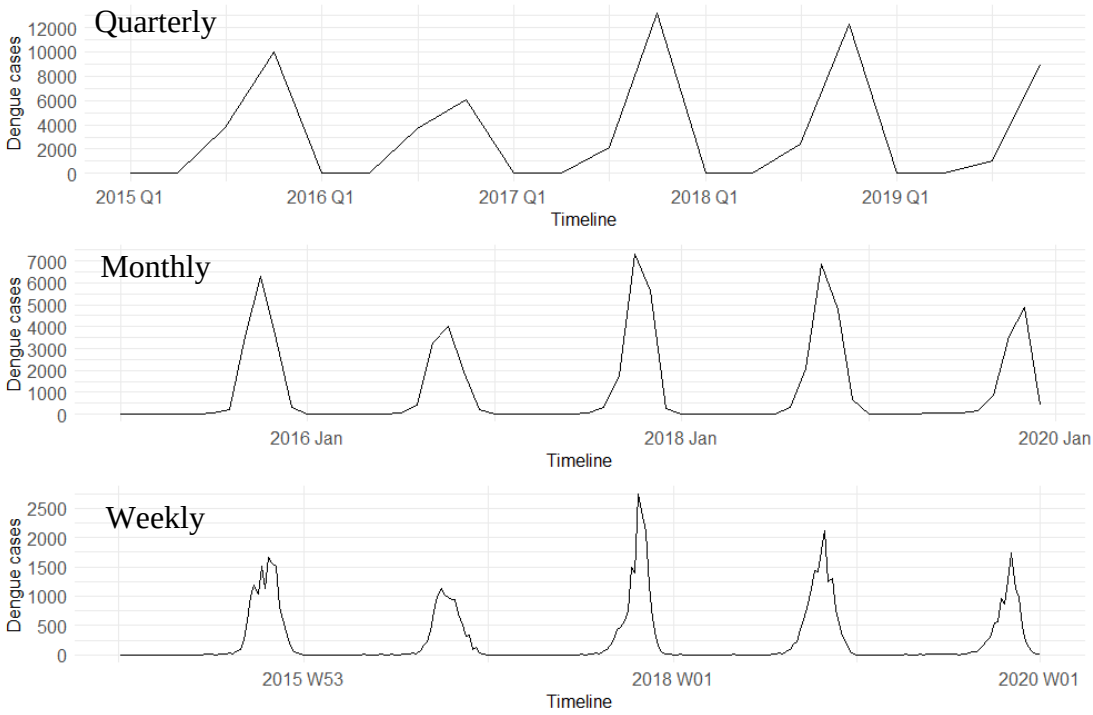


Figure 4.1 Distribution of quarterly, monthly, and weekly occurrence of Dengue in the state

districts. The mean (SD) age was 34.33 (16.78) years with a range from 0 – 120 years (The reported age of two persons was above 97 years).

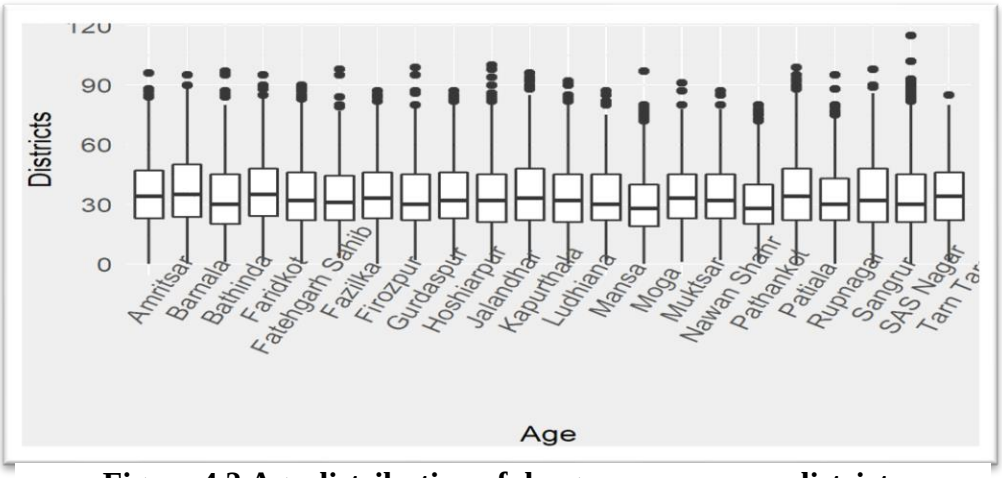


Figure 4.2 Age distribution of dengue cases across districts

Table 4.3 represents the age distribution among males and females. The majority of cases were males (63.94 %). The mean age among females and males were 36 and 34 years, respectively. The maximum number of cases were in the age group of 25-39 years (32% in both males and females), followed by 40-59 years. The mean age reported across districts varied from 32 (Pathankot) to 39 years (Barnala) among females and 29 (Moga) to 36 years (Faridkot) among males.

Table 4.3 Age group-wise distribution of Dengue among females and males

Characteristic	Female, N = 22,717[1]	Male, N = 40,289[1]	p-value[2]
Age (Mean, SD)	36 (17)	34 (17)	<0.001
Age groups			<0.001
< 1 year	79 (0.3%)	254 (0.6%)	
Under 5 years	224 (1.0%)	344 (0.9%)	
5 - 24 years	5,996 (26%)	13,309 (33%)	
25 - 39 years	7,333 (32%)	12,804 (32%)	
40 - 59 years	6,617 (29%)	9,932 (25%)	
> 60 years	2,468 (11%)	3,646 (9.0%)	

[1] n (%)
[2]Welch Two Sample t-test; Pearson's Chi-squared test

4.2.2 Exploration of Climatic and Environmental factors

The minimum and maximum mean land surface temperature (night) recorded across the state were 4.08 degrees Celsius from 20th to 27th December 2019 and

28.43-degree Celsius from 13 to 20th July 2019, respectively. The average minimum and maximum land surface temperature (day) recorded across the state were 12.7 degrees Celsius from the 20th to 27th of December 2019 and 46.4 degrees Celsius from 18 to 25 May 2019. The average daily cumulative precipitation recorded in the state was 101.7 mm and ranged from zero to 4,842.06 mm (22nd September 2018).

The average NDVI recorded in the state was 0.54 ranging from 0.16 (01-15 July 2017) to 0.83 (16-29 February 2016). The state's mean elevation is 198.8 meters, ranging from 132.7 to 513.9 meters. The mean slope of sub-districts in the state was 2.6 and ranged from1.5 to 10.6. The sub-districts in the state varied from purely rural/ town areas to 56.7% urban areas. The percentage of spatial grid cells with a built-up area at a 50% threshold in a sub-district varied from zero to 32.5%.

4.2.3 Spatial distribution of Dengue.

Figure 4.3 represents the annual dengue incidence across districts. It varied from 4.75 per lakh (Fazilka district in 2016) to 210 per lakh (Sahibzada Ajit Singh Nagar district in 2017), with a median annual incidence of 36 cases per lakh population. The choropleth maps suggested that the dengue incidence has remained higher in the southeastern districts of the state for the majority of the years during the study period. Further, the dengue incidence across the western border districts is increasing over the years.

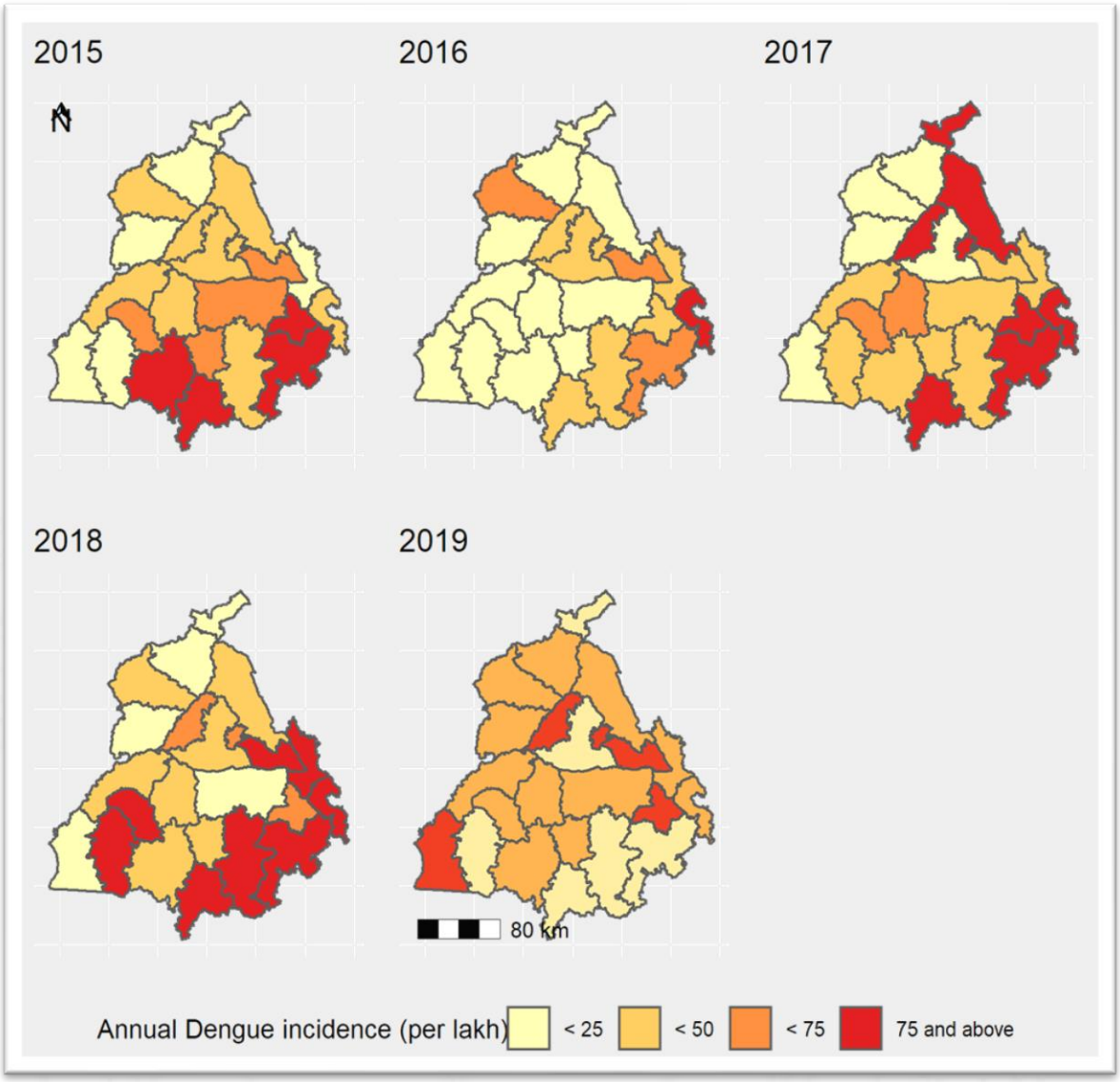


Figure 4.3 Annual Dengue incidence among districts

4.2.4 Spatial distribution of socio-demographic factors

4.2.4.1 Population density

Figure 4.4 represents the spatial distribution of population density estimates. The mean population density across sub-districts in the state was 599 persons per sq. km, ranging from 248 to 4,959 persons per sq. km**.** The least densely populated sub-districts were Dhar Kalan, Lambi, Sangat, and Bhunga (248, 269, 273, and 282 persons per sq. km. Verka, Ludhiana-I, Jalandhar east, and Pathankot sub-districts were the most densely populated (4,959, 4,751, 2,532, and 1,870 persons per sq. km).

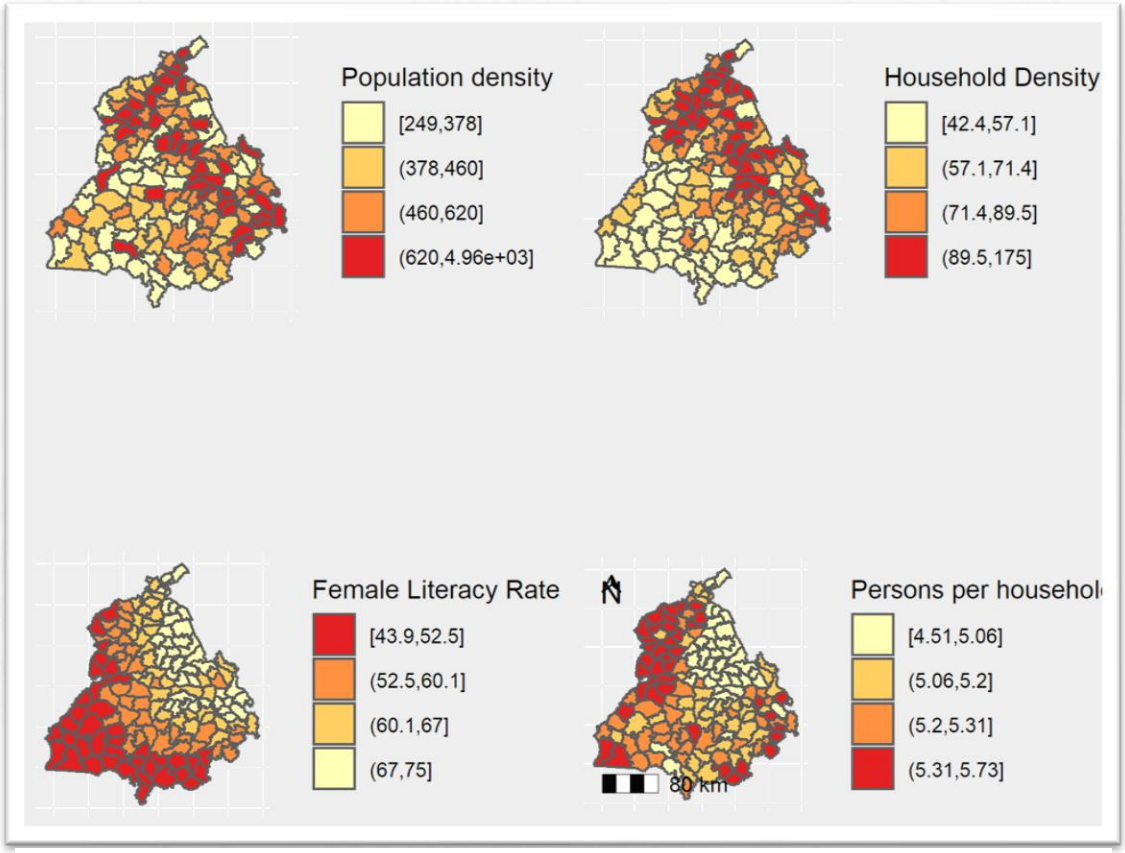


Figure 4.4 Spatial distribution of socio-demographic factors

4.2.4.2 Household density

Figure 4.4 represents the spatial distribution of household density. The mean household density across sub-districts was 74 households per sq. km, ranging from 42 to 175 households per sq. km. The sub-districts with the lowest household density were Makhu, Muktsar, and Sultanpur Lodhi (42, 43, and 44 households per sq. km), and the sub-districts with the highest household density were Pathankot, Gharota, and Sujanpur(175, 144, and 139 households per sq. km).

4.2.4.3 Female literacy rate

Figure 4.4 represents the spatial distribution of the female literacy rate. The mean female literacy rate across sub-districts was 59.4%, ranging from 43.9 to 74.9%. The sub-districts with the lowest female literacy rates were Lehra Ganga, Valtoha, and Jhunir (43.9, 44.1, and 44.6%, respectively and the sub-districts with the highest female literacy rates were Talwara, Bhunga, and Hajipur (74.9, 73.5, and 72.8% respectively.

4.2.4.4 Persons per household

Figure 4.4 represents the spatial distribution of average persons per household. The average number of persons per household across sub-districts was 5.1, ranging from 4.5 to 5.7 persons per household. The lowest number of persons per household was in Adampur, Tanda, and Talwara sub-districts (4.5, 4.5, and 4.6, respectively), and the highest were in Valtoha, Gandiwind Tatla, and Andana sub-districts (5.7, 5.7, and 5.6 respectively).

4.2.5 Spatial distribution of environmental factors.

4.2.5.1 Elevation and slope.

Figure 4.5 represents the distribution of mean elevation and slope across sub-districts. Northeastern bordering sub-districts had the highest elevation. The top five subdistricts with the highest mean elevation were Dhar Kalan (513.9 m), Talwara (433.7 m), Bhunga (322.3), Majri (312.1 m), and Sujanpur (302.2 m). The southwestern subdistricts were low lying with the lowest elevation levels in Guru Har Sahai (143.2 m), Arnival (142.42 m), Abhohar (140.9 m), Jalalabad (138.9 m), and Khuian Sarwar (136.3 m). The slope was highest in the northeastern region and declined towards the southern region.

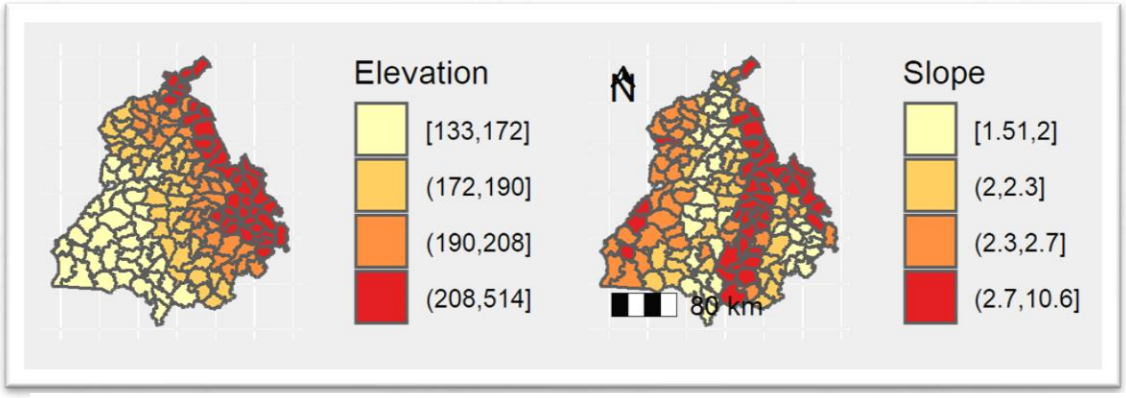


Figure 4.5 Spatial distribution of elevation and slope across sub-districts.

4.2.5.2 Urbanization and built-up area

Figure 4.6 represents the spatial distribution of urbanization and built-up area across sub-districts. The Sub-districts as urban pockets in a district were found across the state; however, the density of these urban pockets was higher in the southern and southeastern regions. The total number of sub-districts which lacked urban areas were 16 (10.67%), namely Bamial, Bhunarheri, Chohla Sahib, Gandiwind Tatla, Jhunir, Kalanaur, Khadur Sahib, Mahal Kalan, Mamdot, Narot Jaimal Singh, Naushera Pannuan, Saroya, Sherpur, Sidhwanbet, Tarsikka, and Aarniwal. The highest urban density was in the Verka sub-district (56.7%), followed by Ludhiana-I (45.5%) and Jalandhar -East (25.5%). The distribution of built-up areas was similar to urbanization. The lowest built-up area was in Bhunarheri, Ghanaur, Talwara, and Dhar Kalan sub-districts, and the highest was in Ludhiana-I (32.5%), followed by Verka (23.8%) and Jalandhar East (14.13%).

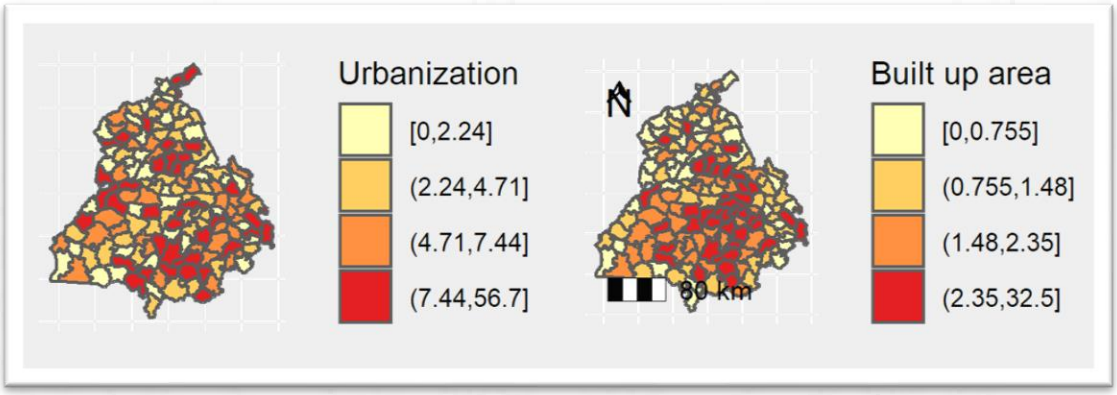


Figure 4.6 Spatial distribution of urbanization and built-up across sub-districts.

4.2.6 Time series features of Dengue.

Figure 4.7 represents the time series Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) plots. Significant positive time series

autocorrelation was present at a lag of four quarters, one month, and up to six weeks. Additionally, a negative autocorrelation was present at a lag of two quarters, 3-7 months and 10-42 weeks. The PACF plots suggested a significant positive autocorrelation of the residuals (after removing the effects explained by the previous intermediate lag period) at one month and one-week lag period. The Hurst coefficient of quarterly, monthly, and weekly time series of dengue occurrence in the state was 0.5, 0.91, and 0.99, and the spectral entropy measure was 0.28, 0.53, and 0.72, respectively.

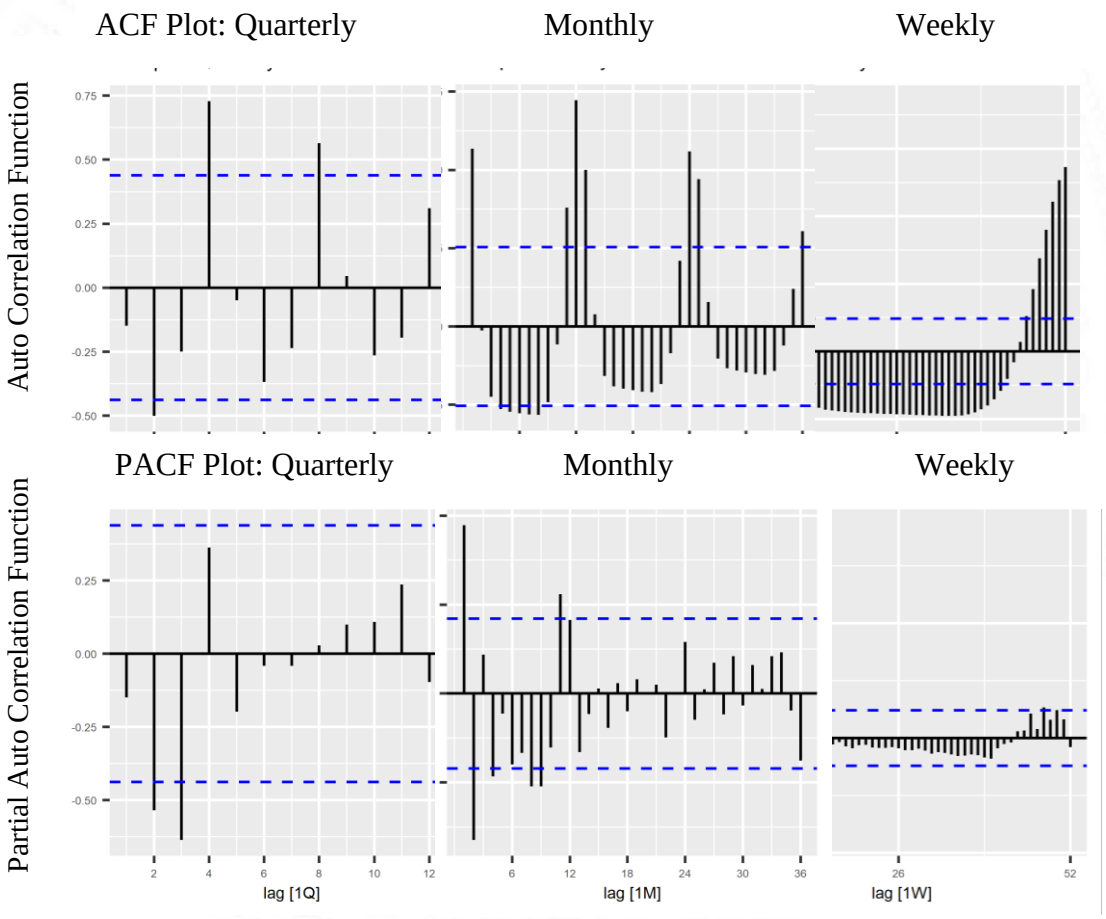


Figure 4.7 Autocorrelation of dengue occurrence in the state

Figure 4.8 represents the time series decomposition of monthly dengue occurrence into seasonality, trend, and remainder components. The measures for the strength of seasonality for quarterly, monthly, and weekly timestamps were 0.93, 0.91, and 0.85, respectively, with a seasonal peak at ten months. Similarly, trend strengths were 0.31, 0.14, and 0.07, respectively, with a seasonal peak at 45 weeks. The remainder component did not show significant patterns and seemed white noise.

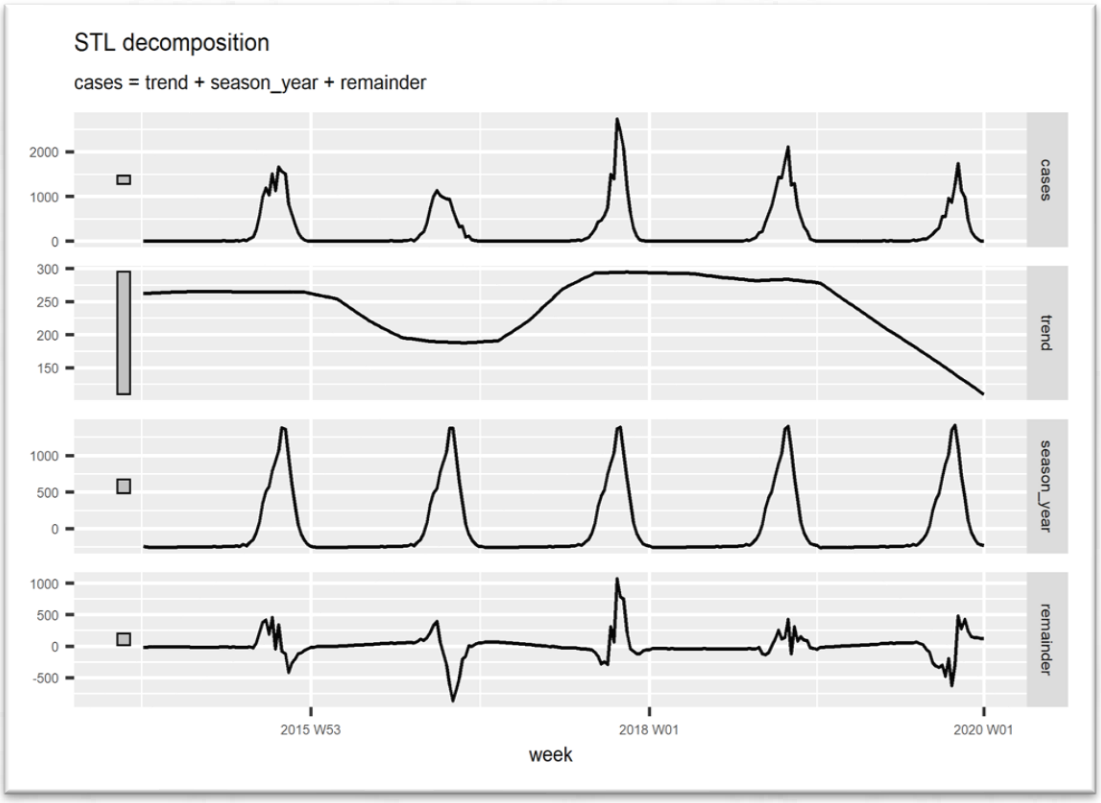


Figure 4.8 Time series decomposition of monthly dengue occurrence in the state

4.2.7 Spatio-temporal epidemiology of Dengue.

Figure 4.9 represents the space-time distribution of dengue incidence at monthly intervals across sub-districts. The space-time distribution plot highlights seasonality with darker shades depicting high-incidence months. Though the overall

dengue incidence was lower among sub-districts in the northern region, the incidence is increasing over the years.

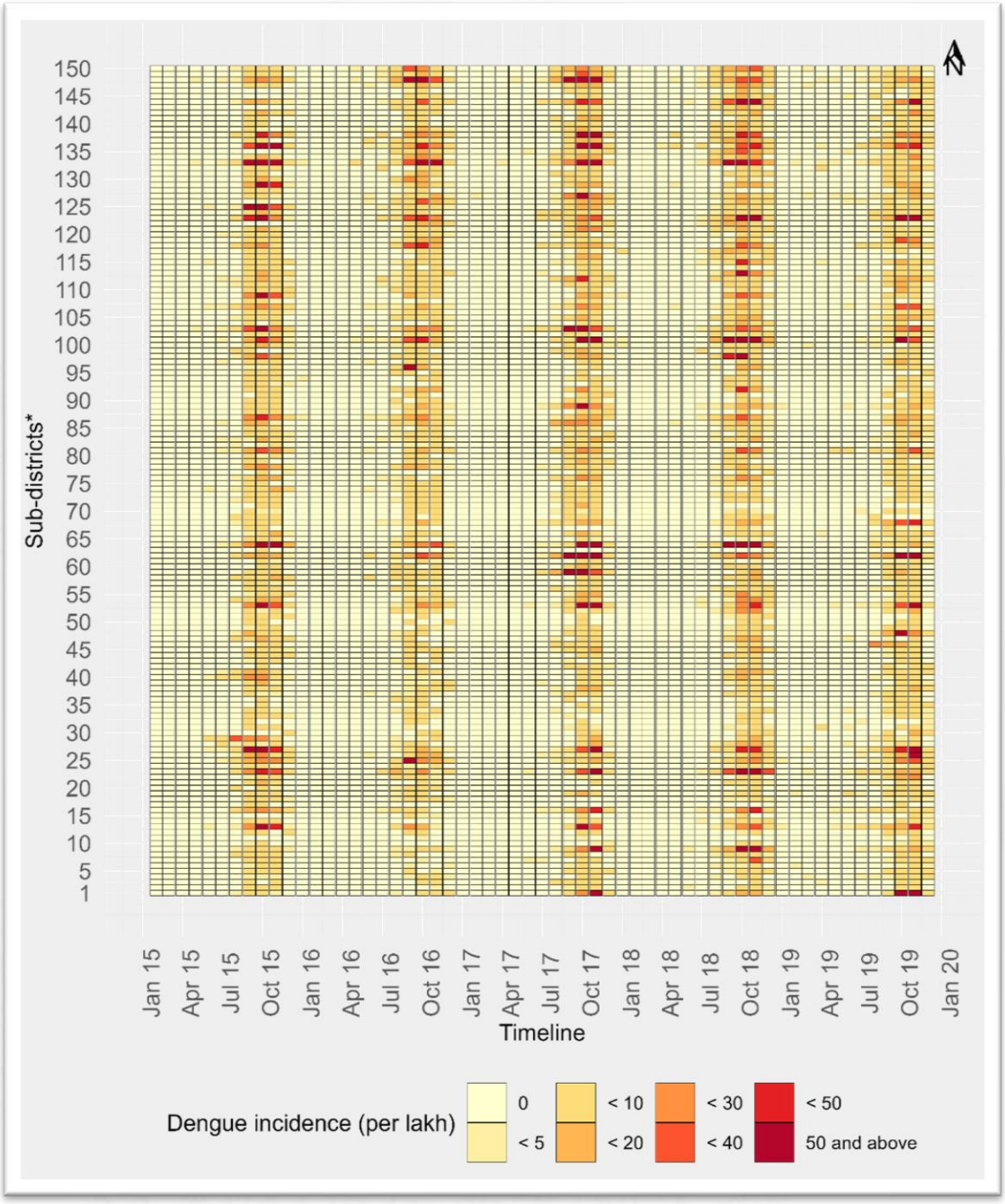


Figure 4.9 Hovemoller diagram: Space-time distribution of dengue across subdistricts.

The X axis depicts timeline; Y axis represents spatial features; and the color represents the dengue incidence at a given time and place. * The 150 subdistricts in the state are arranged from North to South on the Y axis from top to bottom

.

4.2.8 Spatio-temporal Disease risk mapping of Dengue.

Figure 4.10 represents the spatiotemporal distribution of monthly crude standardized incidence rates across subdistricts. The spatiotemporal choropleth grid map depicts the seasonality and presence of Dengue across the state during high-incidence months. Further, it provides additional evidence of increasing incidence in the northern region with time. Also, a shift towards a perennial pattern of Dengue from seasonal occurrence is observed among higher incidences in southern and southeast regions. Sub-districts with SIR 1.5 and above for at least ten months during the study period included Sirhind, Bassi Pathanan, Mansa, Nawan Shahr, Kotkapura, Patiala, Aur, Rupnagar, and Sangrur.

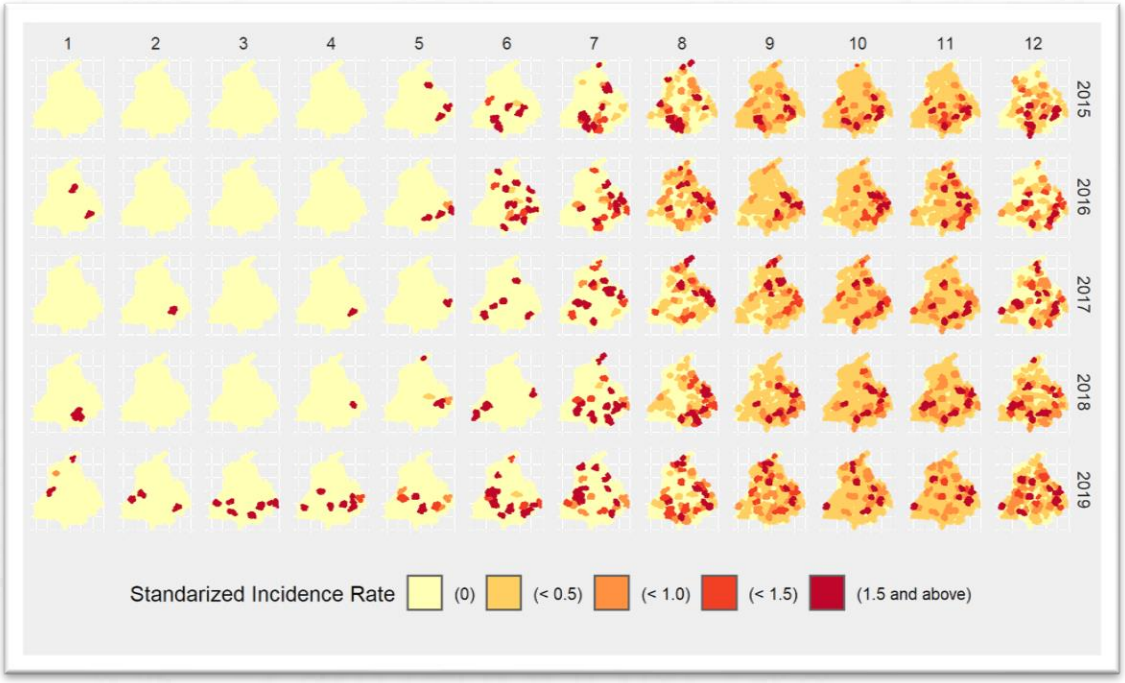


Figure 4.10 Spatiotemporal distribution of Standardized Incidence Rates among sub-districts

4.2.9 Spatio-temporal distribution of climatic and environmental factors

4.2.9.1 Temperature

Figure 4.11 represents the monthly time series plot of land surface temperatures and choropleth map of average temperatures at the sub-district level**.** A seasonal pattern with the highest average monthly temperatures in April-July and lowest in January-February and December was observed. For monthly timestamps, the minimum and maximum monthly average land surface temperature (night) across the state recorded was 7.20 degrees Celsius (December 2019) and 26.38-degree Celsius (July 2019), and LST (day) was 16.05 degree Celsius (December 2019) and 42.04

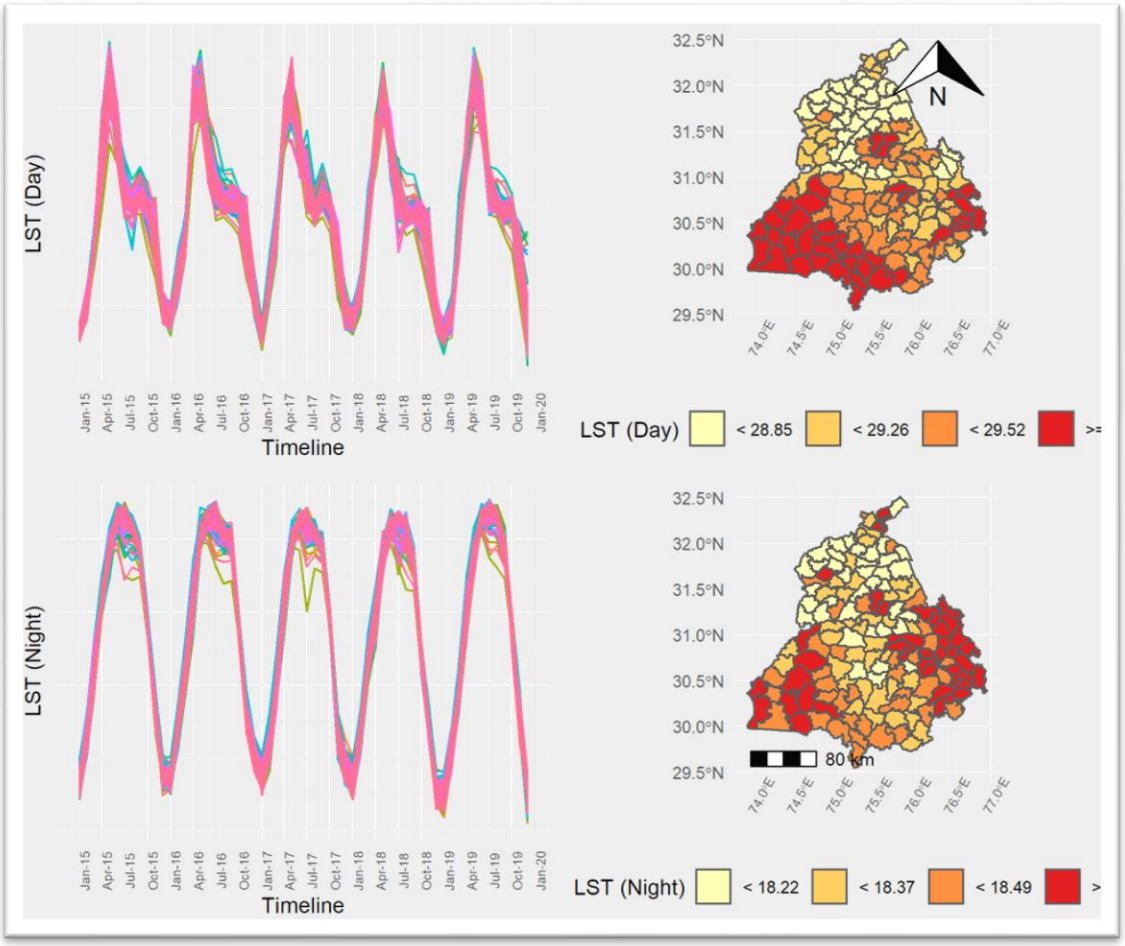


Figure 4.11 Space time distribution of land surface temperatures across subdistricts.

degree Celsius (May 2018) respectively. On STL decomposition analysis of average monthly land surface temperatures in the state during night and day, seasonal strength of 0.99 and 0.98 and trend strength of 0.32 and 0.20, respectively, was observed. The southwestern and southeastern parts of the state had higher average temperatures during the study period. No significant temperature trends were observed within sub-districts (Kendall trend test p-value across sub-districts varied from 0.61 to 0.99 and 0.28 to 0.87 for day and night temperature, respectively).

4.2.9.2 Rainfall

Figure 4.12 represents the monthly time series plot and choropleth map of cumulative precipitation per sq. km across sub-districts. A seasonal pattern was observed with maximum rainfall in the months of June-September. For monthly timestamps, the average cumulative rainfall in the state was 3096.74 mm and ranged from zero (October 2019) to 12,139 mm (July 2017). On STL decomposition, a seasonal and trend strengths of 0.80 and 0.11 were observed. Southwestern sub-districts recorded lower cumulative rainfall per sq. km. No significant rainfall trend

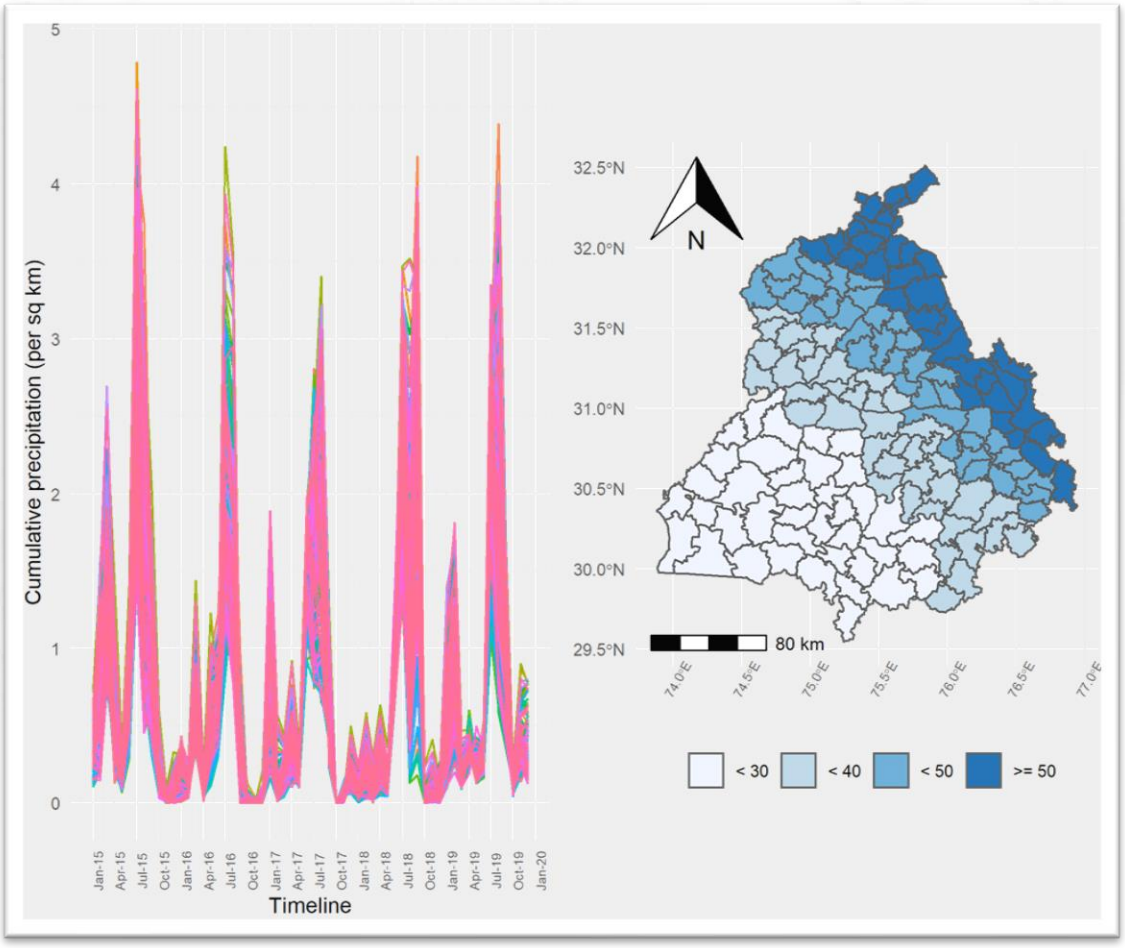


Figure 4.12 Space time distribution of cumulative rainfall (mm per sq. km) across sub-districts

was observed within sub-districts (Kendall trend test p-value across sub-districts varied from 0.45 to 0.99).

4.2.9.3 Humidity

Figure 4.13 represents the monthly time series plot and choropleth map of average relative humidity across sub-districts. For monthly timestamps, the average relative humidity ranged from 26.8% (May 2018) to 69.74% (August 2019). On STL decomposition, seasonal and trend strengths of 0.80 and 0.40 were observed. There was a seasonal pattern with the lowest relative humidity in April-June and the highest in July-September. Northeastern and Southeastern subdistricts had higher average

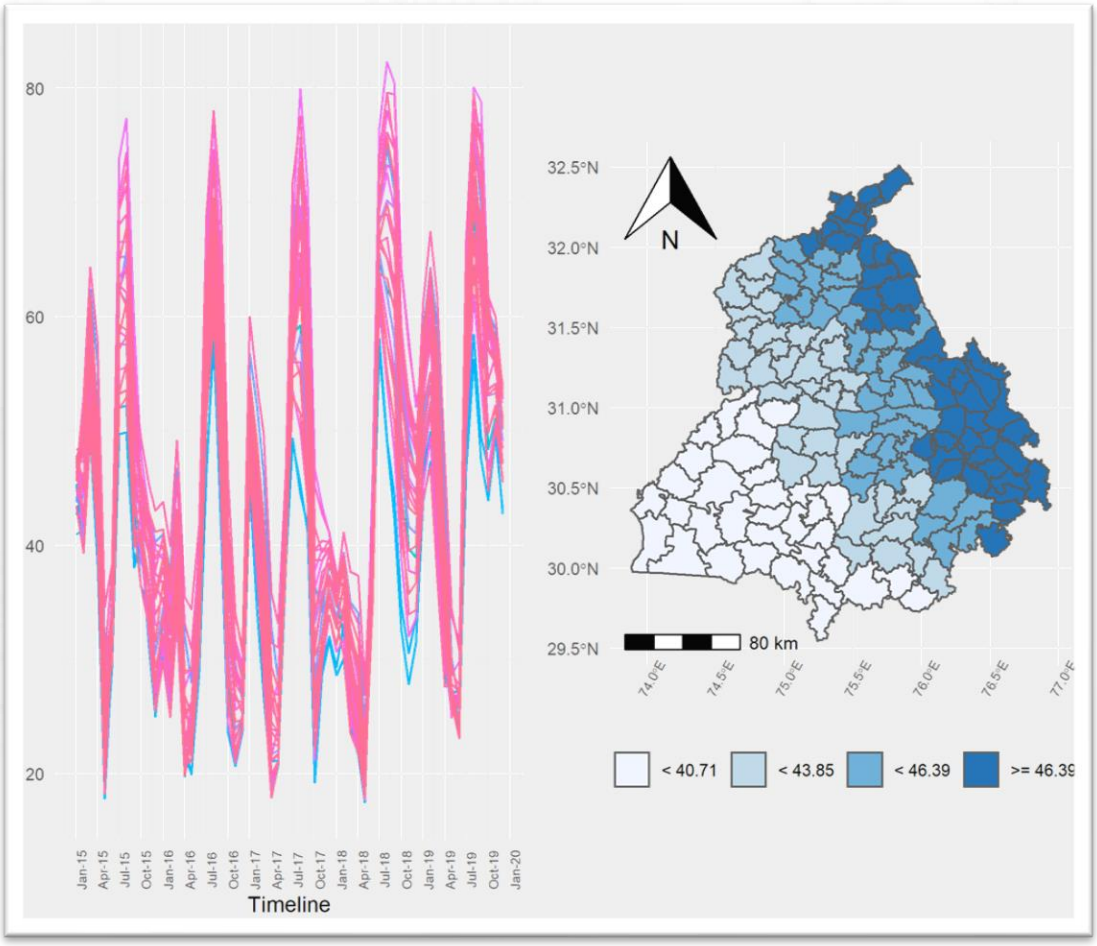


Figure 4.13 Space time distribution of average relative humidity across subdistricts

relative humidity levels than South and Southwestern sub-districts. There was no significant trend of relative humidity across sub-districts (Kendall trend test p-value varied from 0.11 to 0.41).

4.2.9.4 Vegetation

Figure 4.14 represents the monthly time series plot and choropleth map of average NDVI across sub-districts. There was a seasonal pattern with the highest NDVI in January-February (0.7) and the lowest in May-June (0.3). On STL decomposition, seasonal and trend strengths of 0.87 and 0.28 were observed. Southeastern and northern sub-districts had higher average NDVI than southwestern

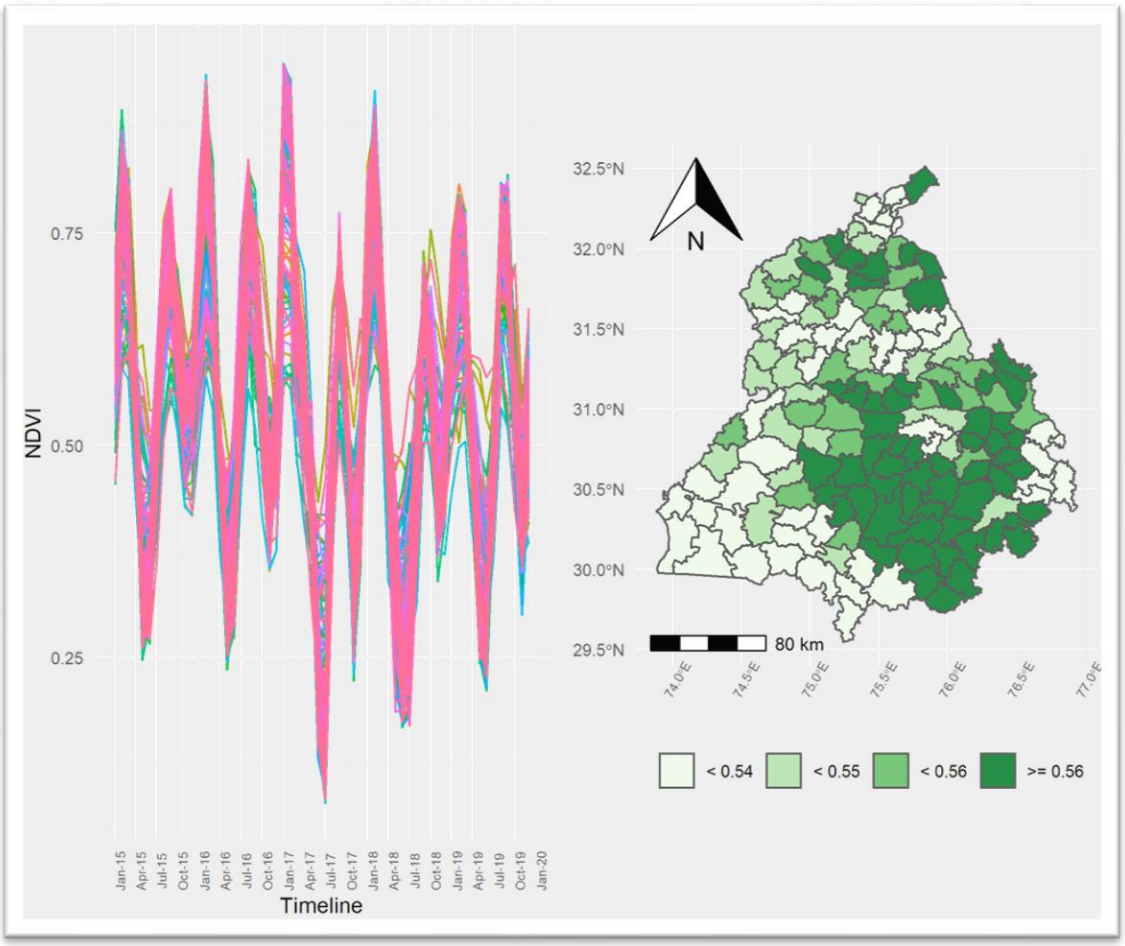


Figure 4.14 Space time distribution of NDVI across sub-districts

sub-districts. No significant NDVI trend was observed (Kendall trend test p-value across sub-districts varied from 0.06 to 0.66).

4.3 Data analysis and Interpretation.

4.3.1 Correlation analysis

4.3.1.1 Association of Dengue with Environmental factors

Figure 4.15 represents the scatter plot matrices for environmental risk factors with dengue incidence. The association between dengue incidence with the level of

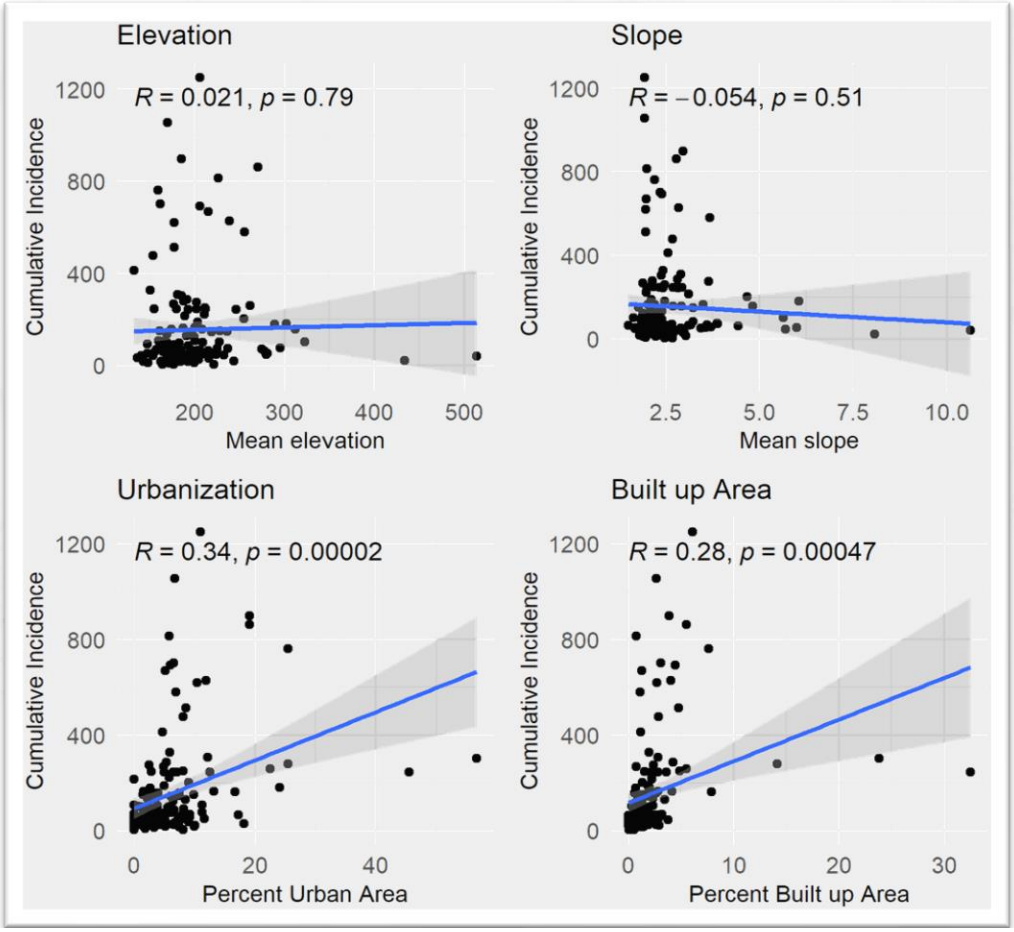


Figure 4.15 Scatterplot matrices: Dengue incidence and environmental factors

urbanization and the built-up area was statistically significant (r = 0.34 and 0.28, p < 0.01).

4.3.1.2 Association of Dengue with Socio-demographic factors

Figure 4.16 represents the scatter plot matrices for the association between dengue incidence and socio-demographic factors. Population density and persons per household were found to be significantly associated with dengue incidence (r = 0.23 and -0.17, respectively, p < 0.01)

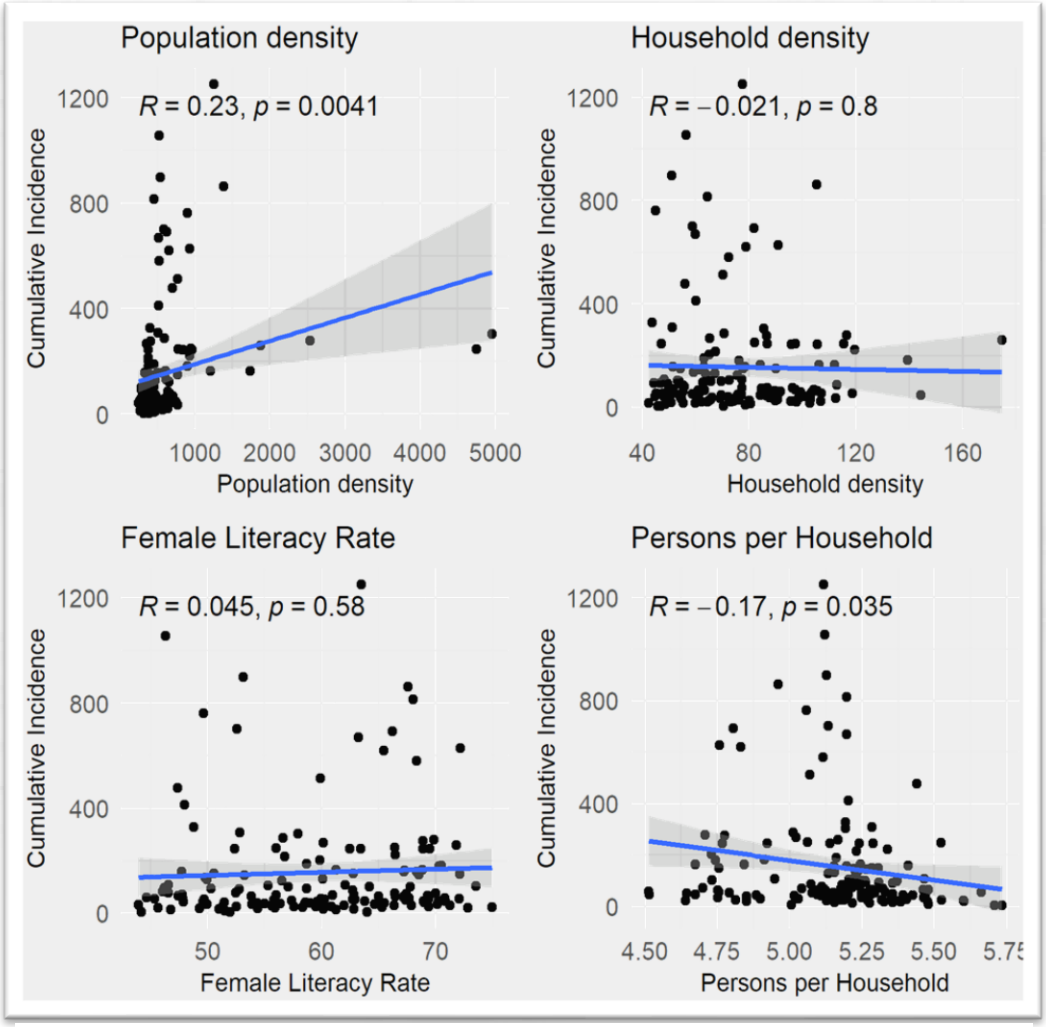


Figure 4.16 Scatter plot matrices: Dengue incidence and socio-demographic factors

4.3.1.3 Association of Dengue with climatic variables

4.3.1.4 Correlation within temperature variables from multiple sources.

Figure 4.17 represents the correlation matrix between land surface temperatures and additional climatic risk factors. The correlation coefficients of land surface temperature (night) were statistically significant with land surface temperature (day), minimum temperature at 2 meters, maximum temperature at 2 meters, WET bulb temperature, minimum temperature at 10 meters, maximum temperature at 10 meters, and dew point temperature (r = 0.80, 0.98, 0.91, 0.90, 0.91, 0.97, 0.90 respectively, $p < 0.05$).

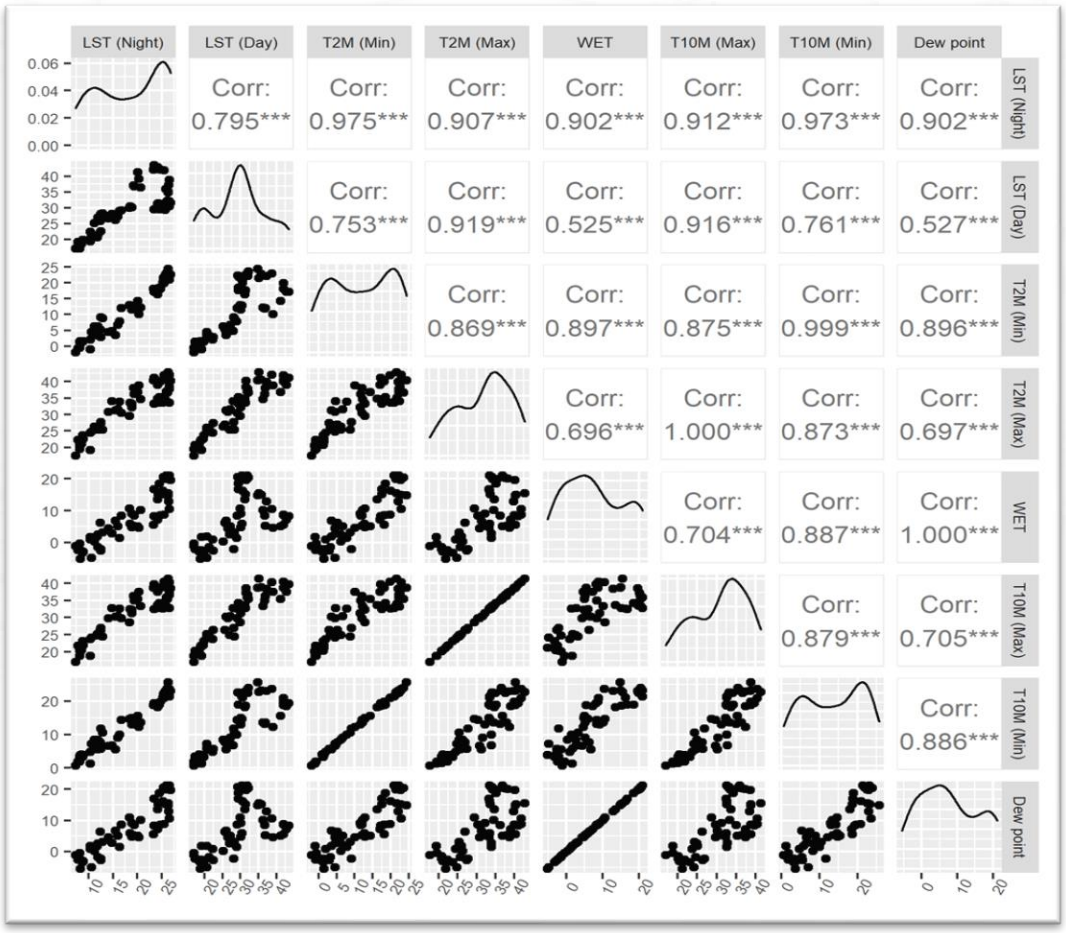


Figure 4.17 Correlation matrix: Temperature

4.3.1.5 Correlation within environmental variables.

Figure 4.18 represents the correlation matrix between environmental variables. The correlation between elevation and slope was statistically significant ($r = 0.76$, $p<0.05$). Similarly, the correlation between urbanization and built-up area was statistically significant ($r = 0.8$, $p<0.05$).

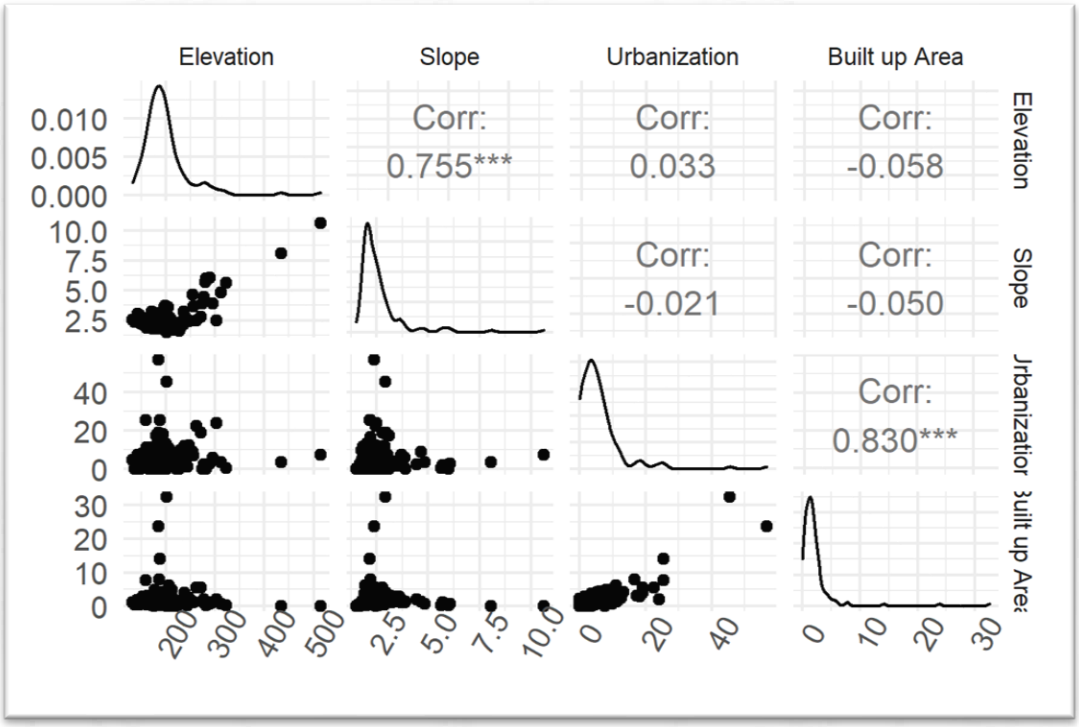


Figure 4.18 Correlation matrix: Environmental factors

4.3.1.6 Correlation within socio-demographic factors

Figure 4.19 represents the correlation matrix between socio-demographic factors. The correlation of population density with household density and the female literacy rate was statistically significant ($r = 0.43$ and 0.23, respectively, $p < 0.05$).

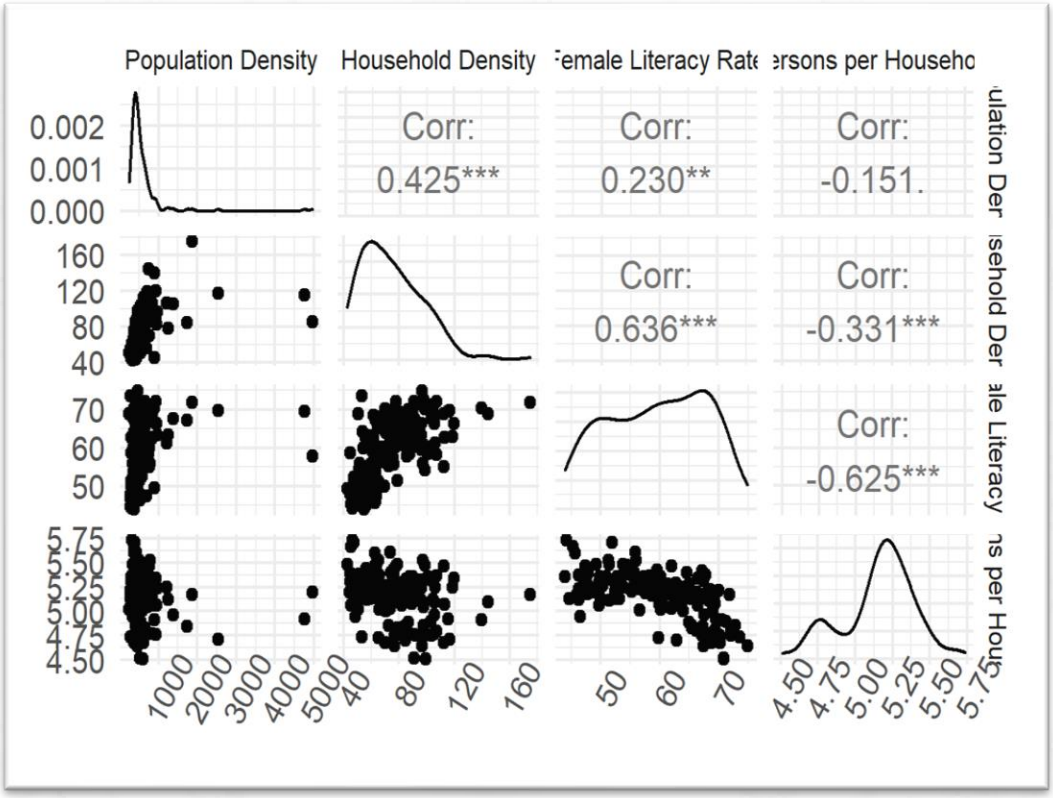


Figure 4.19 Correlation matrix: Sociodemographic variables

4.3.2 Spatial auto-correlation analysis

Figure 4.20 represents the neighbourhood matrix at the sub-district level. It had 150 areal units with a mean of 5.28 neighbours. The matrix obtained was symmetric, without isolated areal units, and the number of neighbours ranged from 1 to 9. The subdistricts with the maximum neighbours were Batala (Gurdaspur district) and Ludhiana (West) (Ludhiana district). The subdistricts with minimum neighbours were Bamial (the northernmost block in Pathankot district) and Sardulgarh (the southernmost block in Mansa district).

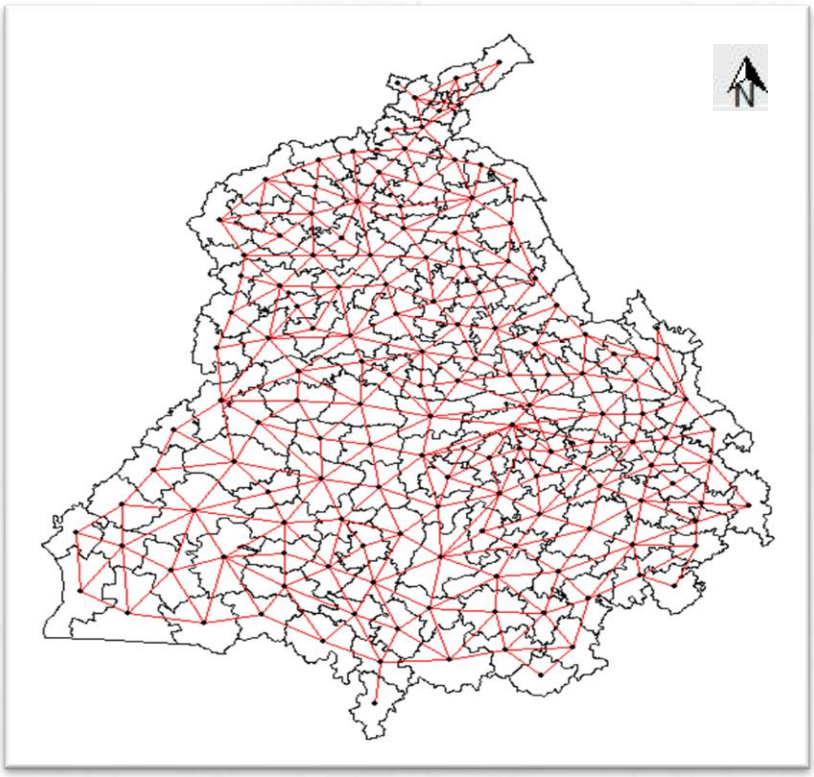


Figure 4.20 Neighborhood matrix

4.3.2.1 *Global estimates of spatial clustering*

Table 4.4 represents *Moran's I's* test statistic and p-values at multiple timestamps. *Moran's I* was statistically significant, with a positive value suggesting spatial clustering among sub-districts for multiple periods ($p<0.05$). The expected Moran's value for the defined neighbourhood and spatial weights was -0.00671. *Moran's I* was statistically significant for annual timestamps for 2016 – 2018. The spatial clustering of Dengue was most observed in August, followed by July and November.

Table 4.4 Moran's I statistics for spatial clustering of Dengue

Year	Month*	*Moran's I*	Z score	p value
2015	Annual	0.07	1.58	0.06
2016	Annual	0.12	2.58	<0.01
2017	Annual	0.12	2.63	<0.01
2018	Annual	0.08	1.70	0.04
2019	Annual	-0.05	-0.94	0.83
2015	Jun	0.02	2.31	0.01
2015	Jul	0.09	3.56	<0.01
2015	Aug	0.11	4.03	<0.01
2015	Nov	0.08	2.02	0.02
2016	Aug	0.09	2.22	0.01
2016	Sep	0.07	1.71	0.04
2016	Oct	0.16	3.66	<0.01
2016	Nov	0.09	2.08	0.02
2017	Jul	0.10	3.17	<0.01
2017	Aug	0.21	4.72	<0.01
2017	Sep	0.14	3.21	<0.01
2017	Oct	0.13	2.91	<0.01
2017	Nov	0.11	2.46	0.01
2017	Dec	0.11	2.40	0.01
2018	Jan	0.16	5.03	<0.01
2018	May	0.09	5.18	<0.01
2018	Jul	0.09	1.98	0.02
2018	Aug	0.13	2.79	<0.01
2018	Sep	0.09	2.10	0.02
2018	Nov	0.09	2.08	0.02
2018	Dec	0.14	3.5	<0.01
2019	Jul	0.06	1.77	0.04
2019	Aug	0.04	2.28	0.01

* *Moran's I* statistic values for months with significant spatial clustering.

4.3.2.2 Local estimates of spatial clusters

Figure 4.21 represents Local *Moran's I* at a significance level of 0.05 for sub-districts during high-burden months. There were spatial clusters of high-incidence subdistricts with high incidence in the neighbourhood (high-high) and low-incidence subdistricts with low incidence in the neighbourhood (low-low). Also, spatial outliers

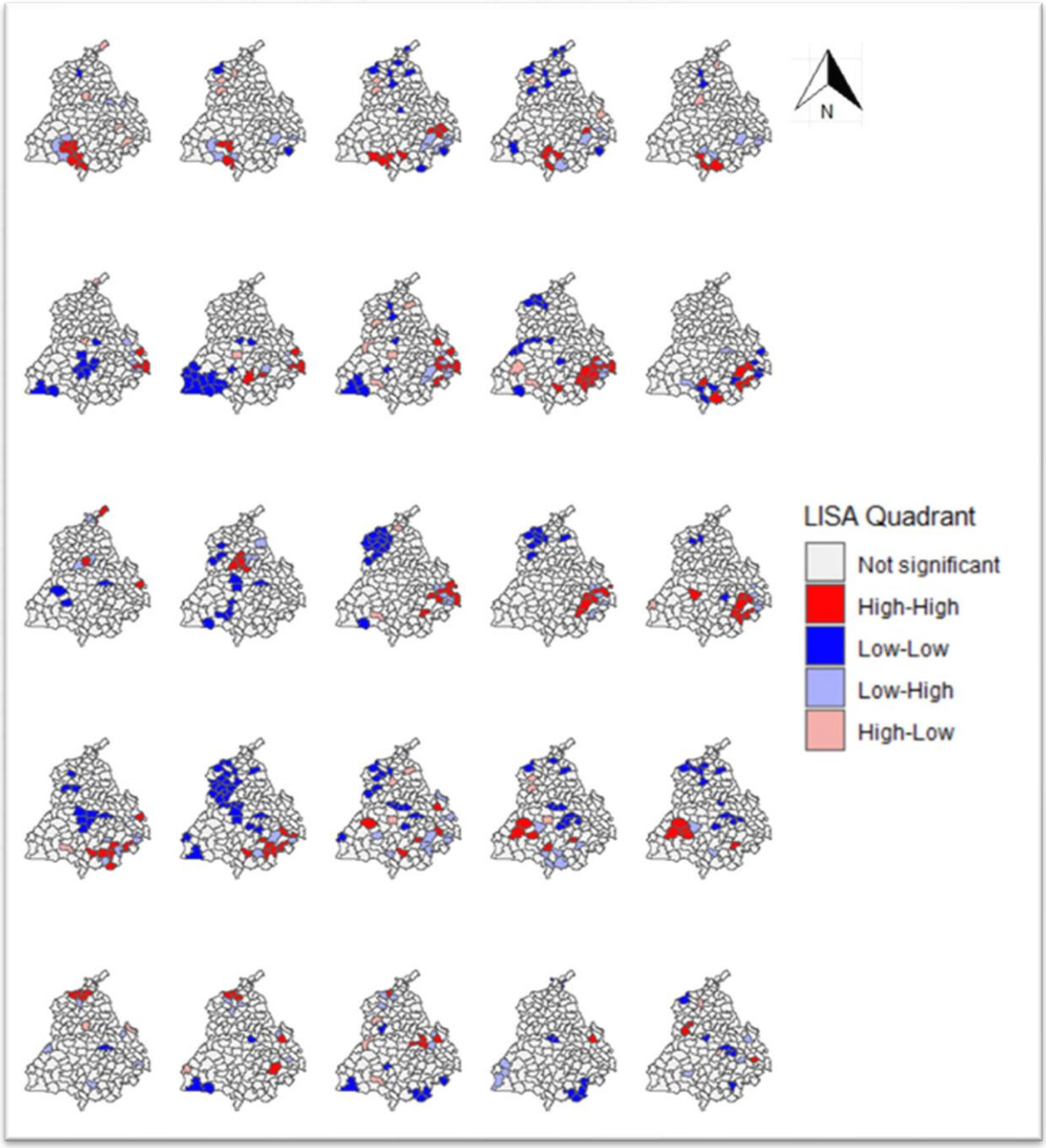


Figure 4.21 Local Moran's for high burden months (p = 0.05)

with a high-low (subdistricts having high incidence but the low incidence in the neighbourhood) and low-high (subdistricts having low incidence despite high incidence in the neighbourhood) were observed. Further, the spatial clusters were dynamic in space and time. Sensitivity analysis at significance levels of 0.01 and 0.1 highlighted the core and spread of spatial neighbourhood clusters. A list of blocks in the respective LISA quadrant and sensitivity analysis results for sub-districts during the high-burden months is provided in appendix F.

4.3.3 Time series cross-correlation analysis

4.3.3.1 Association of Dengue with Climatic factors

Figure 4.22 represents the time series cross-correlation coefficients and scatter plots at specified lags for land surface temperature (night), cumulative precipitation, relative humidity, and NDVI with dengue incidence. The correlation of land surface temperature (night), cumulative precipitation, relative humidity, and NDVI with dengue incidence was positive and statistically significant from a lag of 2 to 5 months (r = 0.52, 0.60, 0.59, and 0.45), 2 to 4 months (r = 0.60, 0.76, and 0.44), 1 to 3 months (r = 0.35, 0.68, and 0.56), and at one month (r = 0.25) respectively ($p < 0.05$). Further,

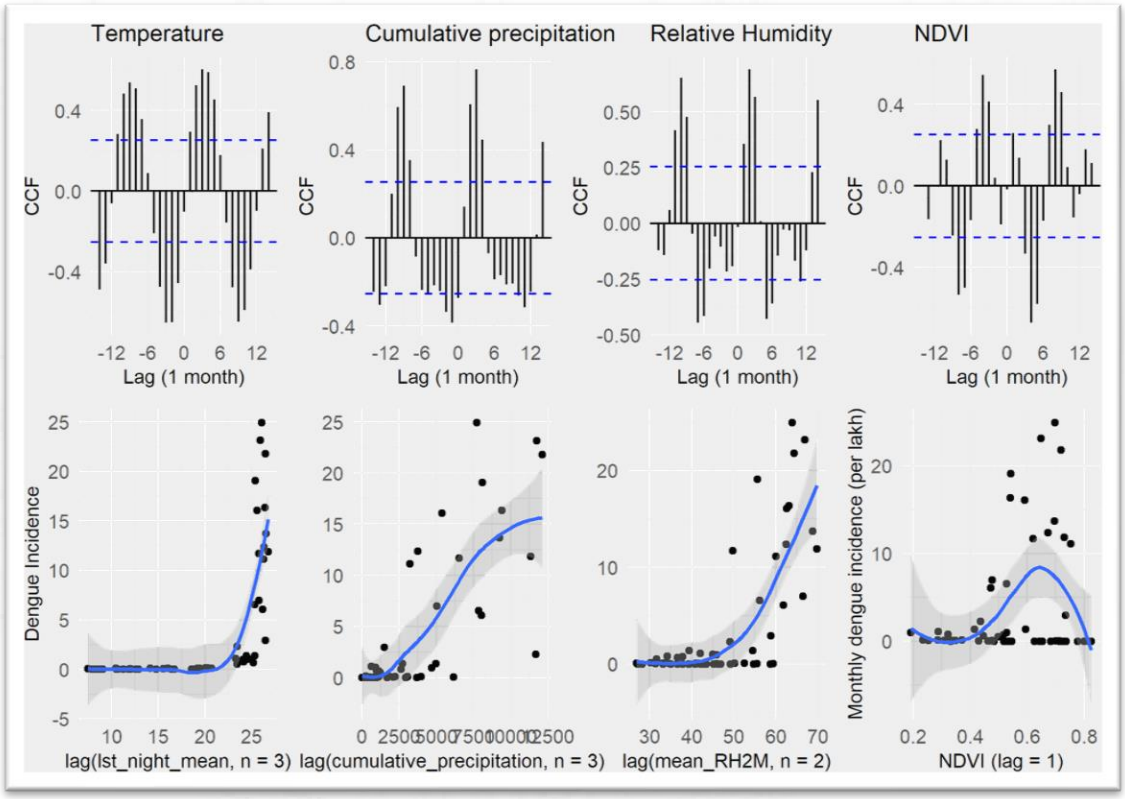


Figure 4.22 Cross correlation coefficient and scatter plots between dengue incidence and climatic factors

the scatter plots suggested non-linear relationships between climatic variables and dengue incidence in the state.

4.3.4 Space-time emerging hotspot analysis

Figure 4.23 represents the spatiotemporal dynamics of dengue incidence rates across subdistricts. Persistent, persistent and intensifying, persistent and diminishing, emerging, oscillating, new, historical, and sporadic hotspot sub-districts were identified in the state (number of sub-districts = 2, 21, 2, 14, 10, 21, 6, and 27, respectively).

Faridkot and Muktsar blocks in the southwestern region, neighbouring each other, were persistent and intensifying hotspots. Sub-districts in SAS Nagar, Fatehgarh Sahib, and Patiala districts in the southeastern region and sub-districts in Faridkot, Moga, Bathinda, Mansa, and Rupnagar districts were persistent hotspots. Fazilka district at the southwestern border and Pathankot at the Northern border were emerging hotspots Mansa, Bathinda, Barnala, Hoshiarpur, and Ludhiana had oscillating

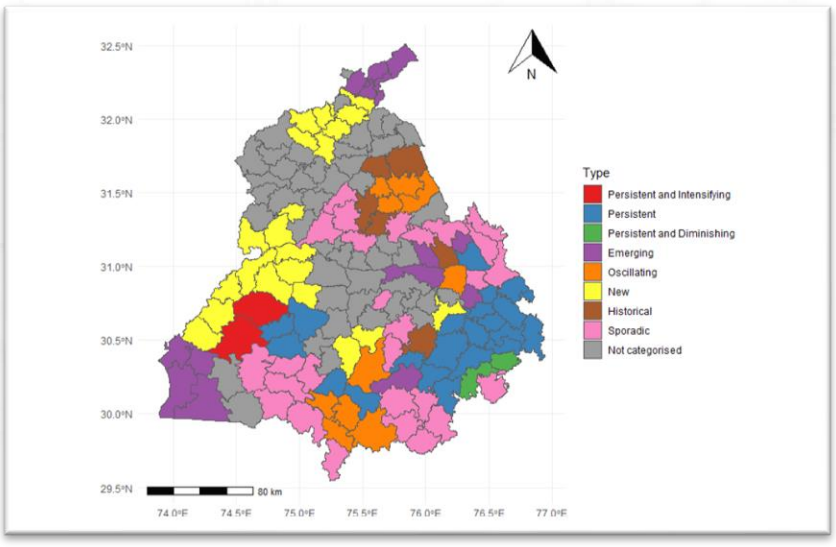


Figure 4.23 Emerging hotspot Analysis

hotspots. New hotspots were observed in districts along the western border, namely Firozpur, Taran taran, and Gurdaspur. Sub-districts with historical hotspots were present in Jalandhar and Hoshiarpur districts.

4.4 Spatiotemporal models

4.4.1 Generalized Linear Models (GLMs)

4.4.1.1 Quasipoisson generalized linear model.

Table 4.5 represents the Quasipoisson Generalized Linear Model estimates for temperature, precipitation, relative humidity, wind speed, and NDVI. The association between climatic variables and NDVI was statistically significant ($p < 0.01$). The dispersion parameter calculated was 8.86, suggesting overdispersion.

Table 4.5 Summary table: Quasipoisson Generalized Linear Model

Variable	**Estimate**	**St. Error**	**Statistic**	**p-value**
Lag3 temperature	0.30	0.01	24.67	< 0.01
Lag3 Precipitation	0.0006	0.00005	12.10	< 0.01
Lag2 Relative Humidity	0.03	0.001	18.51	< 0.01
Lag1 Wind Speed	-1.51	0.07	-19.60	< 0.01
Lag1 NDVI	1.75	0.14	11.88	< 0.01

4.4.1.2 Negative Binomial Generalized Linear Model

The Negative Binomial Generalized Linear Model estimates are represented in **Table 4.6**. The association of lagged climatic factors and NDVI was statistically significant ($p<0.001$). The dispersion parameter, null deviance, and residual deviance were 1.05, 22378.40, and 3419.20, respectively, suggesting adequate goodness of fit.

Table 4.6 Summary table: Negative Binomial Generalized Linear Model

Term	Estimate	Std. error	Statistic	P value
Lag3 Temperature	0.37	0.01	36.34	< 0.001
Lag3 Precipitation	0.0007	0.00009	7.70	< 0.001
Lag2 Relative Humidity	0.04	0.002	15.76	< 0.001
Lag1 Wind Speed	-1.03	0.10	-9.72	< 0.001
lag1_ndvi	0.94	0.18	4.99	< 0.001

Figure 4.24 represents the model diagnostics for the Negative Binomial Generalized Linear model. The top left panel represents the residual plot which does not show significant patterns. The top right panel represents the quantile-quantile (QQ) plot which shows significant deviations from the 45-degree line suggesting non-linearity. The bottom two panels are the plots of the cook's statistics. The bottom left panel represents standardized leverages on the x-axis, cook's statistics on the y-axis, a horizontal dotted line plotted at *8/(n-2p)*, and a vertical dotted line at *2p(n-2p)* where

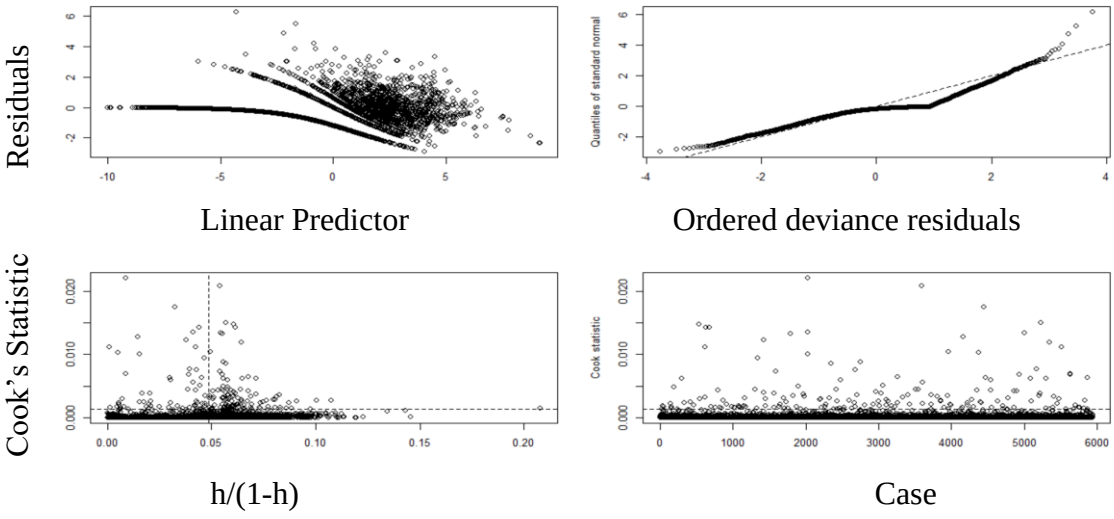


Figure 4.24 Diagnostic plots: Negative Binomial Generalized Regression Model

n is the number of observations and *p* is the numbers of parameters estimated. There are a large number of observations above the horizontal dotted line and to the right of the vertical dotted line suggesting high influence observation points on the model and high leverage compared to the variance of the raw residual at a given point

4.4.2 Generalized Additive Models (GAMs)

Table 4.7 represents the AIC and adjusted r-squared values for initial Generalized Additive Models using multiple lags. The explanatory variables at one month lag were inadequate to provide explainability of relationships between independent smooth variable terms and the outcome (adjusted r squared = -0.79). The lowest AIC value, at this stage, was for the model incorporating a lag of two months for temperature, rainfall, and humidity, one month for wind speed and NDVI, and socio-demographic characteristics.

Table 4.7 Model diagnostics for initial GAM models

Models*	AIC	Adj R squared
T1P1H1	20123.32	-0.79
T2P2H2	18859.82	0.47
T3P3H3	19118.60	0.49
T3P3H2	18993.46	0.47
T3P2H2	18994.60	0.49

*T = Land Surface Temperature (Night), P = Cumulative Precipitation, and H = Relative Humidity at specified lags of 1,2 and 3 months.

Table 4.8 represents model parameters for the T2P2H2 model with nine basis functions for the smooth terms. The model had an AIC of 18,859.82 and an adjusted R squared value of 0.47. The association of smooth terms for temperature, precipitation, relative humidity, wind speed, and NDVI with dengue incidence was

statistically significant ($p<0.01$). The basis dimensionality check showed significant p values for multiple variables indicating the need for an increase in basis functions.

Table 4.8 Model parameters for the T2P2H2 Generalized Additive Model

Term	edf	ref.df	statistic	p.value	K-index
s(lag2_temp)	8.39	8.86	1724.47	< 0.01	0.91
s(lag2_ppt)	6.61	7.14	107.59	< 0.01	0.89^{0}
s(lag2_rel_humidity)	3.87	4.71	676.04	< 0.01	0.78***
s(lag1_wind_speed)	5.87	6.96	237.76	< 0.01	0.77***
s(lag1_ndvi)	6.09	7.19	129.89	< 0.01	0.92*
s(urbanization)	1	1	0.02	0.89	0.89*
s(elevation)	1	1	0.004	0.94	0.89**
s(household density)	1	1	0.03	0.87	0.88**
s(female_literacy_rate)	1	1	0.04	0.84	0.89*
S(area)					0.89

p values for basis dimensionality check: ^{0} = < 0.1; * = <0.05, ** = < 0.01, *** = <0.001

Figure 4.25 represents the partial effect plots for the T2P2H2 model. The plots represent the smooth component functions fitted on the linear predictor scale for the given variable. The visualization of x-axis values for urbanization, elevation, household density, and female literacy rate showed a gap in values, indicating potential inaccuracy of the smooth basis functions in capturing the relationships

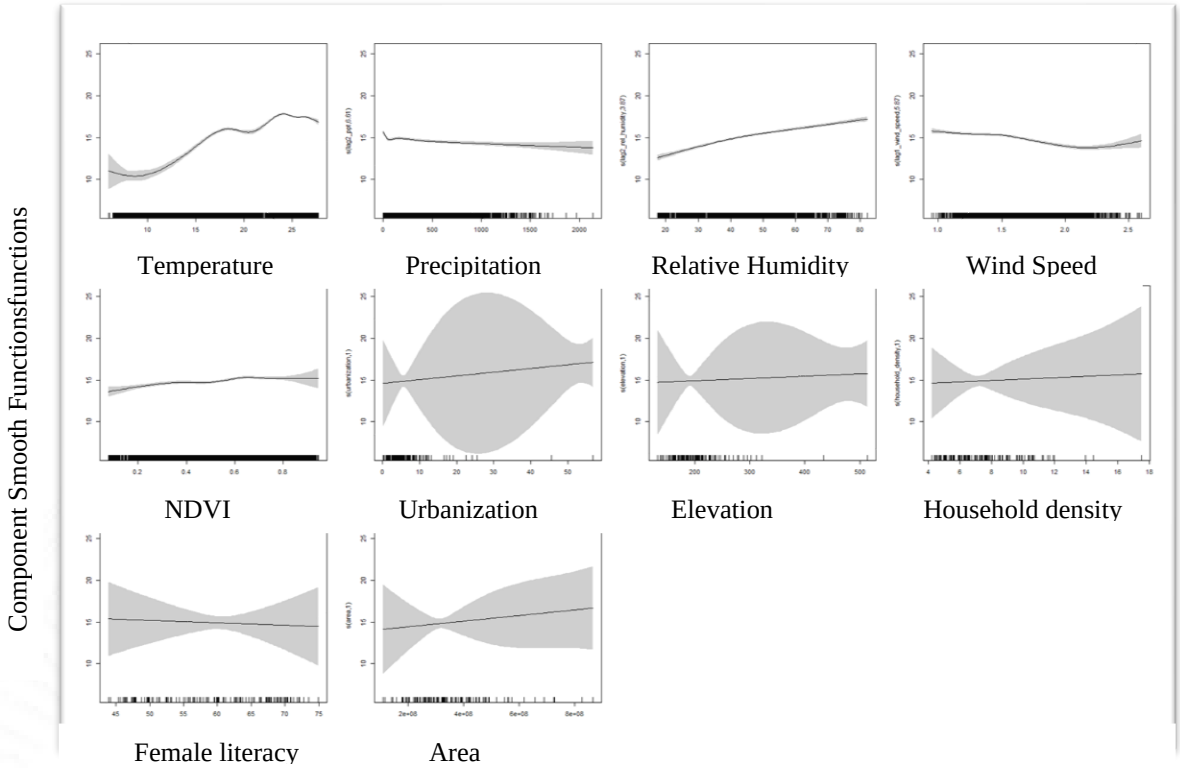


Figure 4.25 Partial effect plots for the best fit initial Generalized Additive Model

Table 4.9 Model estimates: GAMM

Parametric coefficients			
Variable	**Estimate**	**SE**	**P value**
Intercept	-15.6949	0.35672	<0.001
urban_catSemi urban	0.21763	0.2208	0.32
urban_catUrban	0.98342	0.21475	<0.001
elevation_catMid	0.04914	0.25804	0.84
elevation_catHigh	0.77009	0.31543	0.01
literacy_catMedium	-0.02608	0.26099	0.92
literacy_catHigh	0.02242	0.31952	0.94
Smooth terms			
Term	**edf**	**ref.df**	**statistic**	**p.value**
s(NEW_BLOCK)	131.18	142	1981.51	<0.001
s(lag2_temp)	19.36	24.13	294.48	<0.001
s(lag2_range)	4.32	5.59	15.72	<0.01
s(lag2_ppt)	2.01	2.01	2.25	0.33
s(lag2_rel_humidity)	3.27	4.19	10.70	< 0.05
s(lag1_wind_speed)	1.00	1.00	20.00	<0.001
s(lag1_ndvi)	7.90	9.98	61.70	<0.001
s(area)	2.32	2.37	14.52	<0.01
s(month)	5.51	8.00	557.54	<0.001

4.4.3 Generalized Additive Mixed Models (GAMMs)

Table 4.9 represents the model estimates for the Generalized Additive Mixed Model with random effects for the sub-districts. The AIC of the model reduced as compared to the T2P2H2 GAM model to 15,586.98, the adjusted R squared increased to 0.66, and basis functions were adequate.

Figure 4.26 represents diagnostic plots for GAMM. The top left panel is a QQ plot of deviance residuals wherein the majority is on the 45-degree line suggesting uniform distribution. The top right panel represents the residual plot which does not show significant patterns. The bottom left panel represents a histogram of residuals wherein the residuals show near-normal distribution. The bottom right panel does not show a skewed pattern between response and fitted values.

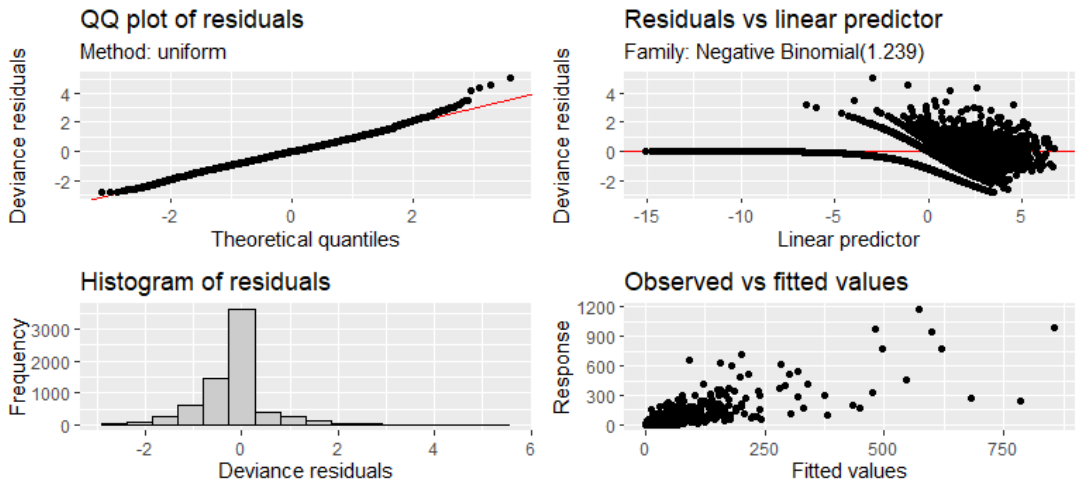


Figure 4.26 Diagnostic plots: Generalized Additive Mixed Model

Table 4.10 represents the *Moran's I* statistic among residuals annually across subdistricts. There was no significant spatial autocorrelation. Similarly, **Figure 4.27**

represents no significant time series autocorrelation among residuals. These findings suggest adequate capture of the spatiotemporal distribution of Dengue by the explanatory variables in the model.

Table 4.10 Residual Spatial Autocorrelation of the GAMM

Year	*Moran's* I statistic	P value
2015	0.02	0.27
2016	-0303	0.71
2017	-0.02	0.66
2018	0.0037	0.41

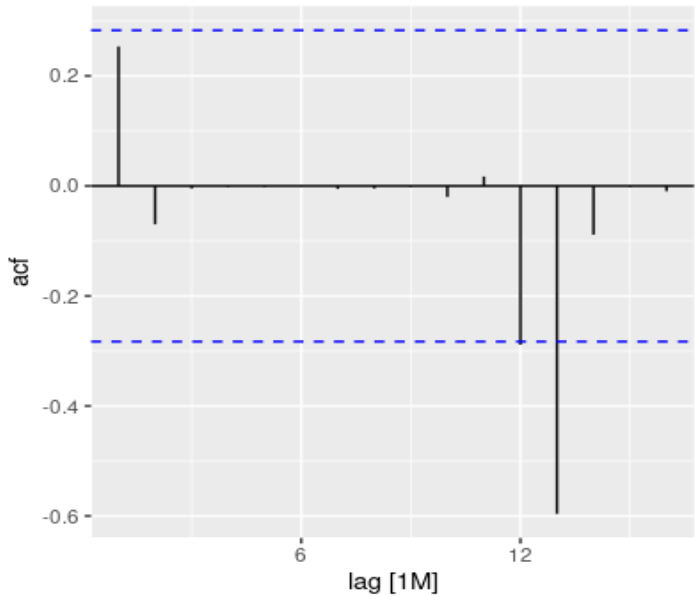


Figure 4.27 Time series Auto-correlation among residuals of GAMM model.

4.5 Dengue forecasting

4.5.1 Forecasts based on Hierarchical Time Series model.

Table 4.11 represents the RMSE values obtained at state, district, and sub-district levels for ARIMA and ETS models with multiple reconciliation methods. The

hierarchical ARIMA model with bottom-up reconciliation method had the lowest observed RMSE (35.94, 117.64, and 803.05 at sub-district, district, and state levels, respectively) and was thus selected for forecasting dengue incidence at state, district, and sub-district levels.

Table 4.11 RMSE values of hierarchical time series forecast models.

Model	State	District	Sub-district
ARIMA Models			
Base	1158.46	117.64	35.94
Bottom Up	803.05	103.92	35.94
Top Down	1158.46	169.73	72.18
Middle Out	1082.10	117.64	50.53
OLS	1150.85	117.18	36.90
MinT	993.07	108.50	37.07
ETS Models			
Base	938.97	119.12	40.05
Bottom Up	917.45	114.67	40.05
Top Down	938.98	108.29	37.67
Middle Out	1034.41	119.13	42.44
OLS	933.7	116.06	40.27
MinT	934.92	112.98	39.18

OLS = Ordinary Least Squares, MinT = Minimum Trace

Figure 4.28 represents the observed and forecasted values for Dengue with 95% CI for 2019 in the state, and **Table 4.12** represents the accuracy levels of the forecast estimates at district and sub-district levels. According to the forecast, there was a steep rise in dengue cases from August-September 2019, which was also observed in the state. During the peak months of Dengue in October-November, the forecast, though predicted the peak, however, was higher and temporally premature as compared to the observed cases.

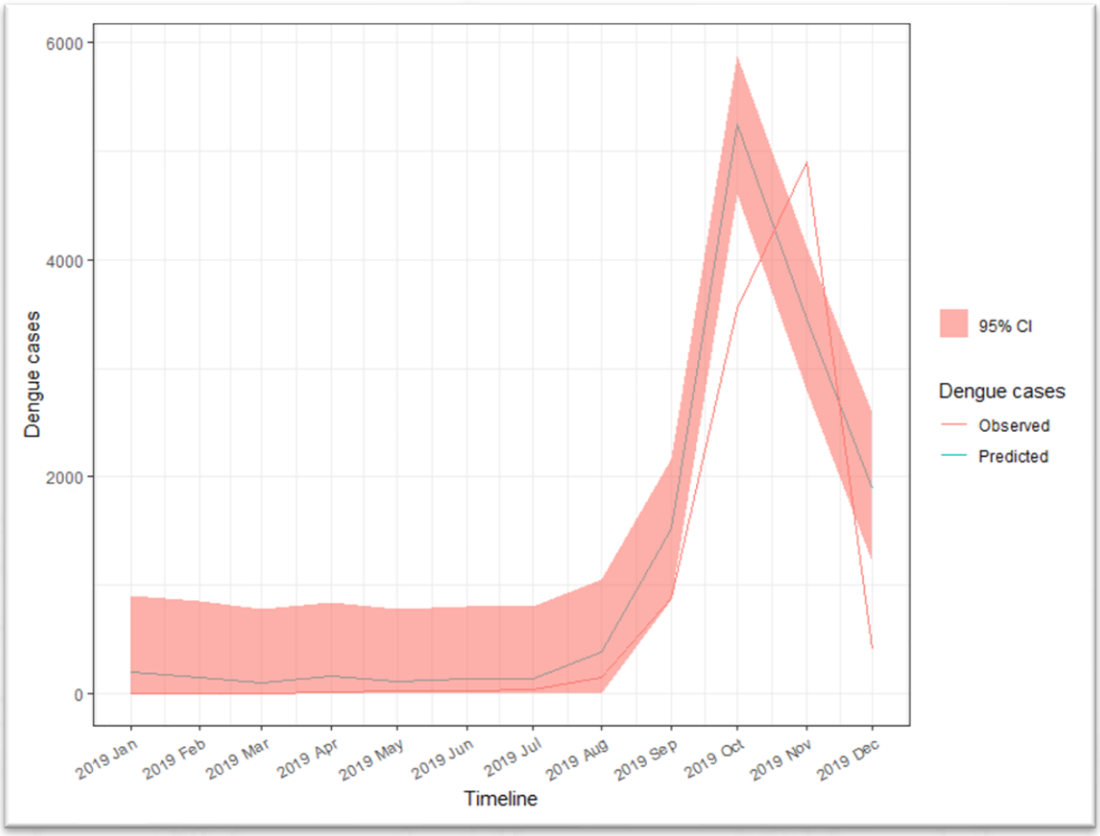


Figure 4.28 Predicted and observed dengue cases in the state using hierarchical time series forecasting model.

Table 4.12 Hierarchical forecast accuracy at district and sub-district levels

Month (2019)	Districts (N = 22) n (%)	Sub-districts (N = 150) n (%)
January	22 (100.00)	148 (98.66)
February	22 (100.00)	150 (100.00)
March	22 (100.00)	150 (100.00)
April	22 (100.00)	149 (99.33)
May	22 (100.00)	150 (100.00)
June	22 (100.00)	150 (100.00)
July	22 (100.00)	150 (100.00)
August	21 (95.45)	148 (98.66)
September	17 (77.27)	129 (86.00)
October	7 (31.81)	78 (52.00)
November	12 (54.54)	94 (62.66)
December	18 (81.81)	127 (84.66)

4.5.2 Forecasts based on Generalized Additive Mixed Model.

The best-fit model of training data from 2015-19 was used for forecasting 2019 dengue occurrence and had RMSE of 790.69, 71.67, and 23.08 at state, district, and sub-district levels. **Figure 4.29** represents the predicted, and observed dengue cases in the state using the GAMM model for forecasting, and **Table 4.13** represents the forecast accuracy at district and sub-district levels.

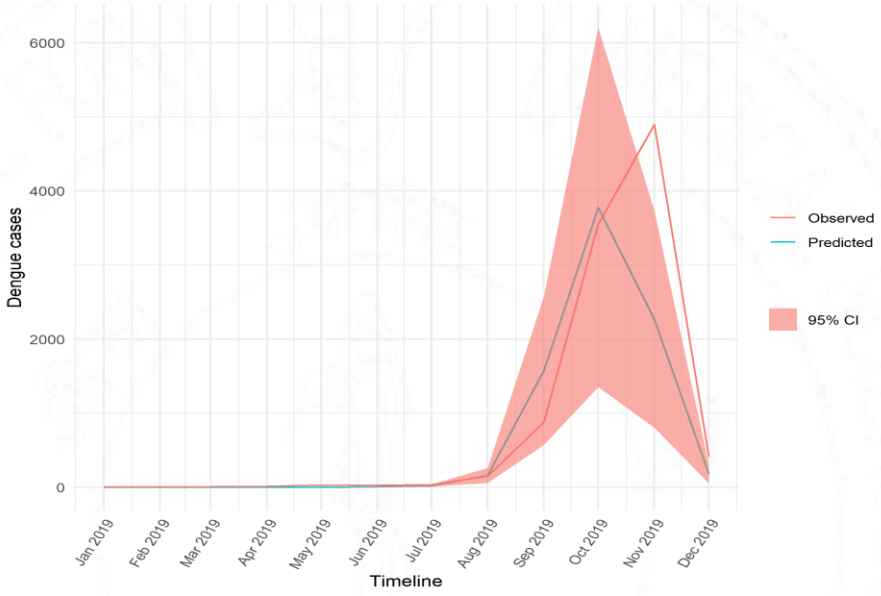


Figure 4.29 Predicted and observed dengue cases in the state using spatio-temporal GAMM forecasting model

Table 4.13 GAMM forecast accuracy at district and sub-district levels

Month (2019)	Districts (N = 22) n (%)	Sub-districts (N = 150) n (%)
January	19 (86.36)	145 (96.66)
February	20 (90.90)	147 (98.00)
March	16 (72.72)	143 (95.33)
April	15 (68.18)	142 (94.66)
May	12 (54.54)	134 (89.33)
June	2 (9.09)	16 (10.66)
July	6 (27.27)	6 (4.00)
August	11 (50.00)	17 (11.33)
September	14 (63.63)	58 (38.66)
October	16 (72.72)	86 (57.33)
November	8 (36.36)	71 (47.33)
December	5 (22.72)	20 (13.33)

CHAPTER 5

DISCUSSION

5 DISCUSSION

5.1 Introduction

This PhD thesis was formulated to understand mechanisms that generate evidence for developing forecasting models using routine dengue data as an empirical dataset. The research project was conceptualized based on the research gaps in using RHIS data from LMICs, especially in the Asian region. The research generated evidence on the Spatio-temporal patterns of dengue in Punjab and its association with climatic, environmental, and sociodemographic factors. Also, forecasting models were developed and evaluated through varied methodological approaches in the previous chapter. The research also contributes to the existing literature by providing reproducible and scalable open-source algorithms that enable data management and analysis of RHIS data. This chapter focuses on the interpretation of the results and potential policy implications for the use of routine data as a decision-support tool in the prevention and control of diseases in public health and suggests recommendations for future RHIS-based research in LMIC settings.

The organization of this chapter is as under: -

1. Summary of the research findings.
2. Policy implications.
3. Strengths and limitations.
4. Future recommendations.

5.2 Summary of the research findings.

5.2.1 Data quality issues in Routine Information Systems.

In the present study, we identified common data quality issues in RHIS data on Dengue line listing, such as lack of standardization in the capture and storage of raw data in separate excel sheets, blank records, duplicates, repeat testing records, missing values, untidy data, and lack of standard formats for age, gender, dates, and addresses. The present study documents a systematic and reproducible logic model for data cleaning of routine health surveillance datasets in LMIC settings. The semi-automated logic algorithm for data extraction was able to extract dates using the *a priori* cleaning codes (automated) for 99.7% of screened values, the algorithm for gender was successful in 97.7% of cases, and the geocoding of standardized addresses was successful in 98.8% records. According to a systematic review that examined the strategies used to address data quality issues in routine health information systems (RHIS) in low- and middle-income countries, it was found that a large proportion of research articles that utilized RHIS data did not provide details on the magnitude of data quality problems or the methods used to mitigate them (Hung et al., 2020)**.** The application of the systematic approach used in this research is similar to systematic techniques used for cleaning RHIS data in previous studies (Gesicho et al., 2020; Maina et al., 2017; Phan et al., 2020); however, the novelty of this research is the application using open source algorithms for RHIS data of a national program in India.

In routine data analytic studies, time series analysis and geostatistical analysis are the two most commonly used analytical methods (Hung et al., 2020). However, the data entry formats for dates in routine health information systems can vary due to the

use of basic data entry platforms such as Microsoft Excel. While this may be sufficient for day-to-day use within the current system, incorporating advancements in data handling and management technologies can lead to a digital transformation of healthcare surveillance. To achieve this, it is essential to engage data producers and users, identify their information needs, build capacity for data use at multiple levels, and strengthen the infrastructure for data use and demand. These measures can enhance the use of data in healthcare systems and improve their overall effectiveness (Nutley and Reynolds, 2013).

In our study, we found that the incidence of missing data values in the extracted analysis-ready data for the selected variables was lower compared to previous studies that used routine datasets. This may be due to the specific variables selected for inclusion in the analysis, which is considered critical for decision-making in the NVBDCP program. The line listing datasets used in the program contain only a limited number of variables deemed essential for the primary data use process. Additionally, we observed that data values that appeared to be missing for a specific variable were more likely to be misplaced within the dataset. Therefore, if alternative columns that could potentially contain relevant information are not examined, the rate of missing data may appear to be higher.

Further, we used satellite imagery datasets for the present study. Though the climatic data were available from meteorological stations, we opted for satellite imagery datasets as it was not found reasonable to interpolate data for the whole state using data from a limited number of locations. This is similar to previously conducted epidemiological studies, wherein missing climatic data were addressed using satellite

imagery data due to the absence of continuous records from meteorological stations. This approach has become widely used by the scientific community due to its low cost and enhanced data availability to the research community (Fuller et al., 2009; Lowe et al., 2011).

5.2.2 Exploratory Data Analysis

The research explored Spatial and time series characteristics of dengue and its association with climatic, environmental, and socio-demographic factors. Exploratory Data Analysis (EDA) is a recently developed statistical approach that emphasizes understanding the capabilities of data before evaluating how well it performs (Bruce and Bruce, 2017; Tukey, 1977). EDA complements confirmatory inferential statistics by minimizing model-building assumption violations and aiding in data comprehension, analysis, and modeling. Exploratory Spatial Data Analysis (ESDA) is a more advanced form of EDA that is utilized for recognizing spatial patterns, creating hypotheses based on spatial characteristics, and identifying appropriate spatial models (De Smith et al., 2018). Similarly, for datasets that incorporate both time and space attributes, Exploratory Spatio-Temporal or Space-Time Data Analysis (ESTDA) is a developing research domain in Geographic Information Science (GIS). Further, the Dengue burden is significantly high in India, accounting for approximately 34% of the global burden (World Health Organization, 2022a). Despite this, the association between dengue and risk factors in the country remains inadequately studied. Also, given the presence of multiple climatic zones and diverse ecological and socio-demographic characteristics within states in India, it is recommended that the

exploration of dengue patterns and their associations be carried out in a local context (Kakarla, Caminade, Mutheneni, Andrew P Morse, et al., 2019).

The current study suggests using monthly intervals for developing forecasting models based on the Hurst coefficient and spectral entropy statistics. Choosing appropriate time stamps is crucial for time series analysis and should consider data availability, long-memory, and noise components (*Forecasting: Principles and Practice (3rd Ed)*, 2021). Also, the selection of time stamps is influenced by various factors such as reporting frequency, data capture mechanisms, and data quality characteristics of health departments (Kumar et al., 2018; Zodpey and Negandhi, 2016; Deeny and Steventon, 2015). In previous dengue modeling studies, daily (Titus Muurlink et al., 2018), weekly (Kakarla, Caminade, Mutheneni, Andrew P Morse, et al., 2019; Phanitchat et al., 2019; Zhang et al., 2019), monthly (Husnina et al., 2019; Jain et al., 2019; Ramadona et al., 2019; Z Xu et al., 2020), and annual (Stolerman et al., 2019) time stamps have been used.

The current proposes an exploratory framework for research projects utilizing routine health information system (RHIS) datasets. This framework allows for a parsimonious approach to exploratory analysis. Such techniques are known to strengthen the plausibility of forecasting models (Strimbu et al., 2017). It is crucial to note that various data analytical approaches have been employed for dengue forecasting in the literature. However, these models are prone to errors if model assumptions and feature selection procedures are inappropriate. Therefore, there is a need for robust statistical exploratory frameworks. The study revealed a significant autocorrelation of dengue incidence and a seasonal component through STL decomposition, suggesting the feasibility of using both approaches with RHIS data.

Additionally, ESTDA highlighted the presence of inter-relationships between risk factors and non-linearity in associations with Dengue, indicating the use of generalized models and efficient feature selection for future analysis and model development. These findings are consistent with previous studies conducted in tropical regions of low- and middle-income countries (Aswi et al., 2019; de Oliveira-Júnior et al., 2019; Swain et al., 2019; Withanage et al., 2018; Zahirul Islam et al., 2018; Zheng et al., 2019).

The present study found that dengue transmission increased in northern hilly sub-districts and exhibited a perennial pattern in the southern sub-districts. The rising temperatures in the elevated northern regions and changing ecology due to increased rainfall in southern regions of the state may have contributed to this trend, as indicated by positive trends on the Seasonal Mann-Kendall test. Although this relationship was not statistically significant in the present study, it is well-known that climate change is affecting disease dynamics globally. A study conducted by 'The Energy and Resources Institute' in India has highlighted the need for inclusive research on climate change and disease dynamics for evidence generation (Dogra et al., 2012). The limited time span of the data collection in the present study may have prevented the establishment of a statistically significant relationship. This limitation is similar to a previous study conducted to understand the climatic associations with dengue occurrence in India (Kakarla, Caminade, Mutheneni, Andrew P Morse, et al., 2019). The lack of larger retrospective dengue RHIS data in the present study can be attributed to recent advances in disease diagnostics and data collection methods (National Centre for Disease Control, Directorate General of Health Services, 2022).

Nevertheless, longer-term studies may be conducted to further evaluate the effects of climate change on dengue occurrence in the region.

5.2.3 Spatio-temporal analysis

This research project showcases the potential of utilizing routine health data for spatiotemporal analysis and the development of open-source algorithms for space-time emerging hotspot analysis in the context of dengue. Routine Health Information Systems capture the place and time of occurrence of a health event/ disease (Jamison et al., 2006). This enables the use of Spatiotemporal methods in understanding disease epidemiology. Integrating data science approaches into Routine Health Information Systems can help health authorities improve dengue prevention strategies and develop targeted public health interventions in identified cluster areas (Garg, 2022). Geographic Information Systems became popular in the United States after the 1970s due to the availability of open data-sharing platforms (Davenhall and Kinabrew, 2022). In India, various initiatives such as the Open Data Initiative by the Government of India (*Open Government Data (OGD) Platform India*, 2022), WHO-IDSP recommendations for increased use of GIS platforms (Directorate General of Health Services, India, 2015), and the National Digital Health Mission (Government of India, 2021) provide opportunities for researchers and administrators to collaborate for understanding disease epidemiology and incorporating advances in spatiotemporal epidemiology for effective resource allocation. These initiatives present a favourable environment for researchers and administrators to work together to better understand and manage dengue and other vector-borne diseases.

This research revealed the emergence of high-low spatial outliers at the beginning of the season in predominantly urban areas surrounded by rural sub-districts, followed by the subsequent spread of high dengue incidence. Additionally, the study observed a dynamic shift of high dengue incidence across sub-districts over time. This phenomenon has been compared to a "forest fire" signature, which is described in a study by Olivier et al. in Delhi, India, where dengue rapidly clustered and spread to adjacent areas beyond the flight range of the mosquito (Telle et al., 2016). These findings further emphasize the importance of integrating GIS into healthcare to develop decision support systems. The results also suggest the need for future research to incorporate non-health sector routine data for a comprehensive understanding of the associated environmental, climatic, socio-demographic, and health system factors for risk-based resource allocation in healthcare, which is a hierarchical system with multiple decision makers (Yan and Haimes, 2011).

The present study created a neighbourhood matrix using Queen's contiguity. It is essential to understand that the selection of the operational definition for creating the neighbourhood matrix should be based on the transmission patterns and epidemiology of the disease under investigation. Dengue, being a mosquito-borne disease, and considering the mobility patterns of the population, the areal units sharing even a single boundary point should be considered. This is in contrast to the approaches for modeling in studies in the domain of one health. Compared to the disease transmission dynamics in plants, zoonoses, and other inter-sectoral areas, human mobility with increased connectivity and commutations for work and leisure, the selection of neighbourhood matrices needs deliberate caution and consideration (O'Sullivan and Unwin, 2014; Pfeiffer et al., 2008).

5.2.4 Dengue forecasting models

In the present study, we used Generalized Additive Mixed Models for forecasting dengue. The best-fit model was developed on training data from 2015-18 based on AIC and model diagnostic plots. The validation of the model was done using testing data from 2019. The model was able to generate forecasts two months in advance with reasonable accuracy. Generalized Additive Models are a class of regression models that allow for flexible modeling of the relationship between the outcome variable and multiple predictors using smoothing functions. In recent years, GAMs have gained popularity in dengue forecasting due to their ability to capture complex nonlinear relationships and interactions between meteorological variables and dengue incidence. A recent study in Malaysia used a GAM to model the relationships between climate variables and dengue incidence. The authors found that temperature, rainfall, and humidity were all significant predictors of dengue incidence, with non-linear relationships between these variables and dengue incidence. The authors used the GAM to develop a dengue forecasting model, which they found was able to predict dengue incidence up to four weeks in advance accurately. (Masrani et al., 2021). Another study used reported dengue data and satellite imagery data from multiple sources for the development of the GAMM forecasting model in Venezuela and found that all the climatic covariates were significantly associated with optimal lagged effect and the smoothed curves captured dynamicity of dengue incidence in the state (Cabrera and Taylor, 2019). Overall, the research was undertaken and the previous studies in the literature demonstrate the potential of GAM models for predicting or forecasting dengue incidence.

We also used hierarchical time series forecasting as an alternate approach for the development of dengue forecasting models. In this research, we applied multiple reconciliation approaches. We found that the hierarchical ARIMA model with bottom-up reconciliation method had the lowest observed RMSE (35.94, 117.64, and 803.05 at sub-district, district, and state levels, respectively) at a one-month advance period. Hierarchical time series forecasting provides a promising approach to improve the accuracy of dengue forecasting by addressing the challenges posed by the hierarchical structure of dengue data. In a study carried out in Sri Lanka using hierarchical time series forecasting, analysis of the forecast accuracy of multiple approaches suggested that the best forecasting approaches for the districts, provinces and the country were not limited to a single approach, and thus evaluation of multiple approaches are recommended (Madushani and Talagala, 2021). Another study from Sri Lanka used a hierarchical time series forecasting method to forecast dengue based on the larval indices and found a significant association with dengue risk at a lag of 1-2 months (Madushani and Talagala, 2021). A recent study from the Thiruvananthapuram district of India looked at dengue risk zones in the district using hierarchical methods and found that around 82.8% of the actual caseload was from the areas modeled as very high and high-risk zones, thus proving the efficacy of such models (Harsha et al., 2022).

5.3 Policy implications

The research undertaken unearthed the spatiotemporal patterns of dengue and its associations with climatic, environmental, and sociodemographic factors for the development of forecasting models up to the sub-district level in the state. Also, we

used open-source, reproducible, and scalable algorithms/frameworks to conduct the study. This research has, therefore, multiple policy implications. The use of interdisciplinary approaches by linking data from non-health sectors and GIS is recommended by WHO-India Joint Monitoring Mission to strengthen disease surveillance.(Directorate General of Health Services, India, 2015). This study is unique in that it is the first to utilize data from the RHIS in conjunction with satellite imagery and census data to analyze the epidemiology of dengue fever in the state. By utilizing open-source platforms and reproducible algorithms, the study aligns with the open data policies of both the World Health Organization and the National Data Policy (National Data Sharing and Accessibility Policy | Department Of Science & Technology, 2022). These methods provide scalable algorithms for future research in low- and middle-income countries, allowing for the development of routine data-driven models.

The current study found that the state had a high annual incidence of dengue ranging from 33.6 to 52.0 per 100,000 population, and the incidence showed seasonality. This finding is consistent with our previous study that investigated the decadal trends of dengue across all states in the country, which showed that Punjab had the highest incidence of dengue compared to the national median annual incidence of 6.57 per lakh population (Singh et al., 2022). Thus, the study findings call for an urgent need to undertake proactive measures to reduce the burden of Dengue in the state.

To undertake this research, multiple visits were undertaken, and coordination meetings were carried out with various stakeholders. Though the data was made

available at the state level for this research, there is a need for enhancing data availability and accessibility of RHIS data to academicians and researchers in LMICs for extracting information and knowledge from existing routine data. It has been observed that researchers in low-and-middle-income countries are often faced with challenges in data collection, storage, sharing, and ensuring data quality across sites and sources (Nori-Sarma et al., 2017). Multiple government initiatives have been rolled out in recent decades, such as the Open Government Data platform, National Digital Health Mission, Bhuvan for spatial data, and the Integrated Health Information Portal by the Government of India. Such policy initiatives must be sustained and enhanced further for RHIS data availability for research.

The research findings generated evidence for eco-socio-demographic factors and their associations with Dengue. It was established that the strongest relationship of dengue was present with minimum temperature, cumulative precipitation, relative humidity, and vegetation cover at time lags of around 2-3 months in the state. Such information has policy implications. The information gathered on the relationship between ecological factors such as climate and environmental data, and the occurrence of dengue can inform the planning of prevention and control mechanisms by health departments. However, it is important to note that the lag associations for disease forecasting may vary from place to place and therefore need to be studied in the local context. In India, a study found that temperature and rainfall were most strongly associated with dengue occurrence at lags of 3-8 weeks and 9-20 weeks, respectively (Kakarla, Caminade, Mutheneni, Andrew P Morse, et al., 2019). Meanwhile, a study carried out in Brazil found shorter lag associations of 1-2 months for both temperature and rainfall (Lowe et al., 2018). These variations may be due to the time taken for the

development and reporting of the disease, which is influenced by local ecological and climatic conditions (Farrar and Manson, 2014).

World Health Organization has identified forecasting dengue fever outbreaks as among the main goals to be achieved (World Health Organization, 2012). The present study used GAMM for the development of forecasting models. Such models have an advantage for policy and decision-makers as they are flexible and can be updated with new data. This is important in the context of dengue, where the incidence can vary significantly from year to year and season to season.

Lastly, the use of open-source algorithms in healthcare has significant implications for the future of public health. Maintaining data integrity and consistency has become a challenge with the exponential increase in data volume in the digital universe (Oracle India, 2021). Despite efforts to ensure data quality assurance and control, these principles are not yet widely implemented, particularly in low-and-middle-income countries like India (Smeets et al., 2011). Real-time decision-making in healthcare requires efficient data-cleaning processes to extract useful information from data. However, manual data cleaning processes are time-consuming, resource-intensive, and prone to errors. Developing reproducible algorithms can enable efficient data cleaning, enhance disease burden estimation, and capacity building, and support rapid decision-making, ultimately leading to better disease prevention and control.

5.4 Strengths and limitations

The research undertaken has multiple strengths toward significant contribution in the domain of public health research and inter-sectoral data science applications in understanding disease epidemiology using RHIS. One of the most important strengths is the use of reproducible open-source algorithms, which enable the generation of

research-level datasets from raw RHIS and satellite imagery datasets. The logic algorithms being open-source can be utilized by researchers in the field of public health and epidemiology in future projects on the one hand and can be scaled to other diseases of public health importance in the health care system.

The other important strength is the inclusion of a large dataset collected over a span of five years across a whole state up to the sub-district level in India. This is a first-of-its-kind study from India. Using a statistically robust, systematic approach and peer-reviewed documentation from the initial protocol stage to data cleaning, pre-processing, and analysis adds strength to this research. Also, using established methods in spatial and time series data analytics for ESTDA and futuristic tidy methods for codes/ algorithms provides methodological and scientific strength to the study.

This study could also emphasize the importance and opportunities of an interdisciplinary research. The cost-effectiveness of the data science approach based on existing RHIS and multiple other routine datasets will help researchers to use available resources and add value to the existing pool of RHIS big data in public health. Further, an important revelation from this research is the potential of spatial, time series, and spatiotemporal methods for the development of forecasting models in an iterative and parsimonious manner.

There are a few limitations in this research that might have impacted the study findings. The use of sub-district and district boundaries is based on administrative requirements, and thus the issue of the Modified Unit Area Problem (MUAP) in spatial analysis cannot be ruled out.

Further, because of misreporting, underreporting, and missed cases, routine data sources are often debated for their quality. Though a concern for the present study, the assumptions of space-time randomness make the analysis undertaken valid for research. The patterns found were biologically plausible and in concordance with the literature. Varied levels of detection and reporting rates may have impacted the study findings. Understanding the reasons behind the data anomalies present in routine datasets is critical in guiding interventions to improve data quality. However, its understanding was beyond the scope of the present study.

The algorithm developed in the present study was based on a single disease dataset from the national vector-borne disease control program. Its application in other diseases and program datasets may require additional screening mechanisms on the one hand and may not need some screening steps on the other. Future studies on applying the algorithm for external generalizability will establish the algorithm's robustness for larger use.

Lastly, in this research, data availability on multiple other risk factors, such as entomological, serological, mobility, transportation networks, etc., was a limitation; however, being scalable, algorithms for additional variables as required by health program managers and for research purposes can be incorporated in future studies into the model. Also, the major challenge in the study was related to the quality of the routine health data. We could not undertake analysis beyond the sub-district level nor point-pattern analysis to identify dengue clusters within sub-districts considering quality issues.

5.5 Future recommendations

This thesis could demonstrate the use of RHIS data and its interlinking with routine datasets from multiple sources to understand dengue epidemiology and for the development of forecasting models. Yet, there is a large void and need for future research using RHIS data in LMICs. Future studies with an aim to develop a dashboard-based Early Warning System (EWS) in the state are required, wherein an extension of forecast models prepared in this research can be institutionalized into the public health system of the state. The representation of the disease processes as static figures provides limited insights into their spatial and temporal dynamics. The development of interactive, automated dashboards with adjustable parameters can facilitate a better understanding of disease processes and promote evidence-informed decision-making. Achieving this requires collaboration between healthcare and IT professionals, as such partnerships are increasingly necessary for multiple sectors, including healthcare. We conducted a stakeholder meeting with the health directorate of the state, and the response of the State Project Officer is positive towards such futuristic projects.

Future work with new co-variates, such as entomological, serological, contact rates, and mobility data, among others, should be undertaken to develop more robust and inclusive forecast models. The additional co-variates will also enable an understanding of the prevalent serotypes of dengue in the state and enable a *mileu* for a precision public health approach in the state.

Value addition to existing data using data science also requires parallel work on strengthening the data quality of the ongoing data collection process. Future work to

identify issues, challenges and strategies for data quality improvement is required to be undertaken in the local context for enhanced data quality and for creating a data use culture in the health care system.

Lastly, future work should investigate opportunities for collaboration of academia with on-ground public health professionals. The merger of advancements in information and technology needs to be harnessed toward implementation and action research. Future studies and initiatives towards achieving sustainable goals are to be deliberated in times to come for a greater public health benefit.

6 SUMMARY AND CONCLUSIONS

In the backdrop of ever-increasing data being collected in RHIS, this research project used dengue line listing data from NVBDCP, Punjab, as an empirical dataset and demonstrated its linkages with multiple routine datasets for understanding mechanisms that generate information and knowledge from RHIS.

In this research, the total data extraction and pre-processing files for all variables included in the study were 3,858, with a data volume of 158.8 GB. A total of 66,581 case records were explored, and it was found that the annual dengue incidence in the state varied from 33.4 per lakh to 52.0 per lakh population. Exploration of the patterns revealed a higher dengue incidence in the last quarter of the year, the month of October, and from weeks 41-46. Choropleth maps suggested dynamicity in dengue occurrence across districts. The spatial distribution of risk factors such as population density, household density, urbanization, and built-up area at the sub-district level was patchy within and between districts.

Disease risk mapping highlighted changing epidemiology of Dengue in the state. The dengue incidence in north-western sub-districts was rising, and there was a shift to the perennial pattern in the southern sub-districts of the state. Dengue incidence was correlated with multiple climatic, environmental, and socio-demographic factors. Spatial autocorrelation analysis showed clustering of Dengue in the state. Maximum clustering across sub-districts was present during the onset (July-Aug) and waning (Nov-Dec) of dengue season. The time series cross-correlation analysis showed a significant association of Dengue with climatic and environmental factors at multiple

lags. Space-time emerging hotspot analysis showed that most subdistricts were sporadic hotspots during the study period. Based on findings from the study's data analysis and interpretation phase, we developed spatiotemporal models. The Generalized Additive Mixed Model and hierarchical time series approaches forecasted dengue in the state with reasonable accuracy.

This research has provided value addition to the existing routine health data using open-source, reproducible, and scalable algorithms. To conclude, the study could demonstrate the potential of using RHIS data to understand disease epidemiology and develop forecasting algorithms in resource-constrained settings. Future work should include institutionalizing data science approaches in the health systems and evaluating their potential for preventing and controlling Dengue and other infectious diseases in LMICs.

REFERENCES

REFERENCES

Akter R, Naish S, Gatton M, et al. (2019) Spatial and temporal analysis of dengue infections in Queensland, Australia: Recent trend and perspectives. *PLOS ONE* Moreira LA (ed.) 14(7): e0220134. DOI: 10.1371/journal.pone.0220134.

Ali MA, Ahsan Z, Amin M, et al. (2016) ID-Viewer: a visual analytics architecture for infectious diseases surveillance and response management in Pakistan. *Public Health* 134: 72–85. DOI: 10.1016/j.puhe.2016.01.006.

Andres A (2009) *Measuring Academic Research: How to Undertake a Bibliometric Study*. Elsevier.

Anno S, Hara T, Kai H, et al. (2019) Spatiotemporal dengue fever hotspots associated with climatic factors in Taiwan including outbreak predictions based on machine-learning. *Geospatial Health* 14(2). DOI: 10.4081/gh.2019.771.

Aqil A, Lippeveld T and Hozumi D (2009) PRISM framework: a paradigm shift for designing, strengthening and evaluating routine health information systems. *Health Policy and Planning* 24(3): 217–228. DOI: 10.1093/heapol/czp010.

Asah FN, Nielsen P and Sæbø JI (2017) Challenges for Health Indicators in Developing Countries: Misconceptions and Lack of Population Data. In: Choudrie J, Islam MS, Wahid F, et al. (eds) *Information and Communication Technologies for Development*. IFIP Advances in Information and Communication Technology. Cham: Springer International Publishing, pp. 593–604. DOI: 10.1007/978-3-319-59111-7_48.

Astuti EP, Dhewantara PW, Prasetyowati H, et al. (2019) Paediatric dengue infection in Cirebon, Indonesia: a temporal and spatial analysis of notified dengue incidence to inform surveillance. *Parasites & Vectors* 12(1): 186. DOI: 10.1186/s13071-019-3446-3.

Aswi A, Cramb SM, Moraga P, et al. (2019) Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. *Epidemiology and Infection* 147: e33. DOI: 10.1017/S0950268818002807.

Banu S, Hu W, Hurst C, et al. (2011) Dengue transmission in the Asia-Pacific region: impact of climate change and socio-environmental factors. *Tropical medicine & international health: TM & IH* 16(5). 5: 598–607. DOI: 10.1111/j.1365-3156.2011.02734.x.

Bett B, Grace D, Lee HS, et al. (2019) Spatiotemporal analysis of historical records (2001–2012) on dengue fever in Vietnam and development of a statistical model for forecasting risk. *PLOS ONE* Wen T-H (ed.) 14(11): e0224353. DOI: 10.1371/journal.pone.0224353.

Bouzille G, Poirier C, Campillo-Gimenez B, et al. (2018) Leveraging hospital big data to monitor flu epidemics. *Computer Methods and Programs in Biomedicine* 154: 153–160. DOI: 10.1016/j.cmpb.2017.11.012.

Braga JU, Bressan C, Dalvi APR, et al. (2017) Accuracy of Zika virus disease case definition during simultaneous Dengue and Chikungunya epidemics. *PloS One* 12(6): e0179725. DOI: 10.1371/journal.pone.0179725.

Bruce PC and Bruce A (2017) *Practical Statistics for Data Scientists: 50 Essential Concepts*. First edition. Sebastopol, CA: O'Reilly.

Cabrera M and Taylor G (2019) Modelling spatio-temporal data of dengue fever using generalized additive mixed models. *Spatial and Spatio-temporal Epidemiology* 28: 1–13. DOI: 10.1016/j.sste.2018.11.006.

Chae S, Kwon S and Lee D (2018) Predicting Infectious Disease Using Deep Learning and Big Data. *INTERNATIONAL JOURNAL OF ENVIRONMENTAL RESEARCH AND PUBLIC HEALTH* 15(8). DOI: 10.3390/ijerph15081596.

Choudhri AF, Siddiqui A, Khan NR, et al. (2015) Understanding Bibliometric Parameters and Analysis. *RadioGraphics* 35(3): 736–746. DOI: 10.1148/rg.2015140036.

Churakov M, Villabona-Arenas CJ, Kraemer MUG, et al. (2019) Spatio-temporal dynamics of dengue in Brazil: Seasonal travelling waves and determinants of regional synchrony. *PLOS Neglected Tropical Diseases* Althouse B (ed.) 13(4): e0007012. DOI: 10.1371/journal.pntd.0007012.

Davenhall WF and Kinabrew C (2022) Geographic Information Systems in Health and Human Services. In: Kresse W and Danko D (eds) *Springer Handbook of Geographic Information*. Springer Handbooks. Cham: Springer International Publishing, pp. 781–805. DOI: 10.1007/978-3-030-53125-6_29.

de Oliveira-Júnior JF, Gois G, da Silva EB, et al. (2019) Non-parametric tests and multivariate analysis applied to reported dengue cases in Brazil. *Environmental Monitoring and Assessment* 191(7). 7: 473. DOI: 10.1007/s10661-019-7583-0.

De Smith MJ, Goodchild MF and Longley PA (2018) *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*. Sixth edition. London: Drumlin Security.

Deeny SR and Steventon A (2015) Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. *BMJ Quality & Safety* 24(8). BMJ Publishing Group Ltd: 505–515. DOI: 10.1136/bmjqs-2015-004278.

Dehury R and Chatterjee S (2018) Assessment of health management information system for monitoring of maternal health in Jaleswar Block of Balasore District, Odisha, India. *Indian Journal of Public Health* 62(4): 259. DOI: 10.4103/ijph.IJPH_203_17.

Directorate General of Health Services, India (2015) Joint Monitoring Mission Report. Ministry of Health and family Welfare, Govt of India.

Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India (2021) National Vector Borne Disease Control Programme (NVBDCP). Available at: https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=405&lid=3681 (accessed 30 January 2021).

Dogra N, Srivastava S and Energy and Resources Institute (eds) (2012) *Climate Change and Disease Dynamics in India*. New Delhi: The Energy and Resources Institute.

Etamesor S, Ottih C, Salihu IN, et al. (2018) Data for decision making: using a dashboard to strengthen routine immunisation in Nigeria. *BMJ Global Health* 3(5). BMJ Specialist Journals. DOI: 10.1136/bmjgh-2018-000807.

Fadahunsi KP, Akinlua JT, O'Connor S, et al. (2019) Protocol for a systematic review and qualitative synthesis of information quality frameworks in eHealth. *BMJ Open* 9(3): e024722. DOI: 10.1136/bmjopen-2018-024722.

Farinelli EC, Baquero OS, Stephan C, et al. (2018) Low socioeconomic condition and the risk of dengue fever: A direct relationship. *Acta Tropica* 180: 47–57. DOI: 10.1016/j.actatropica.2018.01.005.

Farrar J and Manson P (eds) (2014) *Manson's Tropical Diseases*. 23. ed. Expertconsult.com. Edinburgh: Elsevier Saunders.

Forecasting: Principles and Practice (3rd Ed) (2021). Available at: https://otexts.com/fpp3/index.html (accessed 30 July 2022).

Fuller DO, Troyo A and Beier JC (2009) El Niño Southern Oscillation and vegetation dynamics as predictors of dengue fever cases in Costa Rica. *Environmental Research Letters* 4(1): 014011. DOI: 10.1088/1748-9326/4/1/014011.

Garg PK (2022) Geospatial Health Data Analytics for Society 5.0. In: Garg Pradeep Kumar, Tripathi NK, Kappas M, et al. (eds) *Geospatial Data Science in Healthcare for Society 5.0*. Disruptive Technologies and Digital Transformations for Society 5.0. Singapore: Springer Singapore, pp. 29–58. DOI: 10.1007/978-981-16-9476-9_2.

Gesicho MB, Were MC and Babic A (2020) Data cleaning process for HIV-indicator data extracted from DHIS2 national reporting system: a case study of Kenya. *BMC Medical Informatics and Decision Making* 20(1): 293. DOI: 10.1186/s12911-020-01315-7.

Glèlè Ahanhanzo Y, Ouedraogo LT, Kpozèhouen A, et al. (2014) Factors associated with data quality in the routine health information system of Benin. *Archives of Public Health* 72(1): 25. DOI: 10.1186/2049-3258-72-25.

Government of India (2021) National Digital Health Mission. Available at: https://ndhm.gov.in/ (accessed 19 February 2021).

Gupta N, Srivastava S, Jain A, et al. (2012) Dengue in India. *The Indian journal of medical research* 136(3). India: 373–390.

Halstead SB (ed.) (2008) *Dengue*. Tropical medicine : science and practice v. 5. London : Hackensack, NJ: Imperial College Press ; Distributed by World Scientific Pub.

Harrison K, Rahimi N and Carolina Danovaro-Holliday M (2020) Factors limiting data quality in the expanded programme on immunization in low and middle-income countries: A scoping review. *Vaccine* 38(30): 4652–4663. DOI: 10.1016/j.vaccine.2020.02.091.

Harsha G, Anish TS, Rajaneesh A, et al. (2022) Dengue risk zone mapping of Thiruvananthapuram district, India: a comparison of the AHP and F-AHP methods. *GeoJournal*. DOI: 10.1007/s10708-022-10757-7.

Hoxha K, Hung YW, Irwin BR, et al. (2020) Understanding the challenges associated with the use of data from routine health information systems in low- and middle-income countries: A systematic review. *Health Information Management Journal*. SAGE Publications Ltd STM: 1833358320928729. DOI: 10.1177/1833358320928729.

Hung YW, Hoxha K, Irwin BR, et al. (2020) Using routine health information data for research in low- and middle-income countries: a systematic review. *BMC Health Services Research* 20(1): 790. DOI: 10.1186/s12913-020-05660-1.

Husnayain A, Fuad A and Lazuardi L (2019) Correlation between Google Trends on dengue fever and national surveillance report in Indonesia. *Global Health Action* 12(1): 1552652. DOI: 10.1080/16549716.2018.1552652.

Husnina Z, Clements ACA and Wangdi K (2019) Forest cover and climate as potential drivers for dengue fever in Sumatra and Kalimantan 2006–2016: a spatiotemporal analysis. *Tropical Medicine & International Health*: tmi.13248. DOI: 10.1111/tmi.13248.

Indian Meteorological Department (2022) Data Supply Portal. Available at: https://dsp.imdpune.gov.in/ (accessed 28 November 2022).

Jain R, Sontisirikit S, Iamsirithaworn S, et al. (2019) Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. *BMC Infectious Diseases* 19(1): 272. DOI: 10.1186/s12879-019-3874-x.

Jamison DT, World Bank and Disease Control Priorities Project (eds) (2006) *Disease Control Priorities in Developing Countries*. 2nd ed. New York : Washington, DC: Oxford University Press ; World Bank.

Kakarla SG, Caminade C, Mutheneni SR, Morse Andrew P., et al. (2019) Lag effect of climatic variables on dengue burden in India. *Epidemiology and Infection* 147: e170. DOI: 10.1017/S0950268819000608.

Kakarla SG, Caminade C, Mutheneni SR, Morse Andrew P, et al. (2019) Lag effect of climatic variables on dengue burden in India. *Epidemiology and Infection* 147: e170. DOI: 10.1017/S0950268819000608.

Kimaro H and Twaakyondo H (2006) Analysing the hindrance to the use of information and technology for improving efficiency of health care delivery system in Tanzania. *Tanzania Journal of Health Research* 7(3): 189–197. DOI: 10.4314/thrb.v7i3.14259.

Kumar M, Gotz D, Nutley T, et al. (2018) Research gaps in routine health information system design barriers to data quality and use in low- and middle-income countries: A literature review. *The International Journal of Health Planning and Management* 33(1): e1–e9. DOI: 10.1002/hpm.2447.

Lambrechts L, Paaijmans KP, Fansiri T, et al. (2011) Impact of daily temperature fluctuations on dengue virus transmission by *Aedes aegypti*. *Proceedings of the National Academy of Sciences* 108(18): 7460–7465. DOI: 10.1073/pnas.1101377108.

Last JM and International Epidemiological Association (eds) (2001) *A Dictionary of Epidemiology*. 4th ed. New York: Oxford University Press.

Latifov MA and Sahay S (2013) Challenges in Moving to "Health Information for Action": An Infrastructural Perspective From a Case Study in Tajikistan. *Information Technology for Development* 19(3): 215–229. DOI: 10.1080/02681102.2012.751575.

Ledien J, Souv K, Leang R, et al. (2019) An algorithm applied to national surveillance data for the early detection of major dengue outbreaks in Cambodia. *PLOS ONE* Lau EH (ed.) 14(2): e0212003. DOI: 10.1371/journal.pone.0212003.

Liu D, Guo S, Zou M, et al. (2019) A dengue fever predicting model based on Baidu search index data and climate data in South China. *PLOS ONE* Samy AM (ed.) 14(12): e0226841. DOI: 10.1371/journal.pone.0226841.

Liu K, Wang T, Yang Z, et al. (2016) Using Baidu Search Index to Predict Dengue Outbreak in China. *Scientific Reports* 6(1). 1: 38040. DOI: 10.1038/srep38040.

Lopez DM, de Mello FL, Dias CMG, et al. (2017) Evaluating the Surveillance System for Spotted Fever in Brazil Using Machine-Learning Techniques. *FRONTIERS IN PUBLIC HEALTH* 5. DOI: 10.3389/fpubh.2017.00323.

Lowe R, Bailey TC, Stephenson DB, et al. (2011) Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. *Computers & Geosciences* 37(3): 371–381. DOI: 10.1016/j.cageo.2010.01.008.

Lowe R, Gasparrini A, Van Meerbeeck CJ, et al. (2018) Nonlinear and delayed impacts of climate on dengue risk in Barbados: A modelling study. *PLoS medicine* 15(7): e1002613. DOI: 10.1371/journal.pmed.1002613.

Lu FS, Hattab MW, Clemente CL, et al. (2019) Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches. *Nature Communications* 10(1): 147. DOI: 10.1038/s41467-018-08082-0.

MacCormack-Gelles B, Lima Neto AS, Sousa GS, et al. (2018) Epidemiological characteristics and determinants of dengue transmission during epidemic and non-epidemic years in Fortaleza, Brazil: 2011-2015. *PLOS Neglected Tropical Diseases* Barker CM (ed.) 12(12): e0006990. DOI: 10.1371/journal.pntd.0006990.

Madushani LS and Talagala TS (2021) Hierarchical Forecasting of Dengue Incidence in Sri Lanka. arXiv. DOI: 10.48550/ARXIV.2112.00992.

Maïga A, Jiwani SS, Mutua MK, et al. (2019) Generating statistics from health facility data: the state of routine health information systems in Eastern and Southern Africa. *BMJ Global Health* 4(5): e001849. DOI: 10.1136/bmjgh-2019-001849.

Maina JK, Macharia PM, Ouma PO, et al. (2017) Coverage of routine reporting on malaria parasitological testing in Kenya, 2015–2016. *Global Health Action* 10(1): 1413266. DOI: 10.1080/16549716.2017.1413266.

Marques-Toledo C de A, Degener CM, Vinhal L, et al. (2017) Dengue prediction by the web: Tweets are a useful tool for estimating and forecasting Dengue at country and city level. *PLoS neglected tropical diseases* 11(7): e0005729. DOI: 10.1371/journal.pntd.0005729.

Masrani AS, Nik Husain NR, Musa KI, et al. (2021) Prediction of Dengue Incidence in the Northeast Malaysia Based on Weather Data Using the Generalized Additive Model. *BioMed Research International* Khani jeihooni A (ed.) 2021: 1–8. DOI: 10.1155/2021/3540964.

MEASURE Evaluation (2022) Routine Health Information Systems. Available at: https://www.measureevaluation.org/our-work/routine-health-information-systems.html (accessed 11 November 2022).

Minale AS and Alemu K (2018) Mapping malaria risk using geographic information systems and remote sensing: The case of Bahir Dar City, Ethiopia. *Geospatial Health* 13(1): 660. DOI: 10.4081/gh.2018.660.

Morin CW, Comrie AC and Ernst K (2013) Climate and Dengue Transmission: Evidence and Implications. *Environmental Health Perspectives* 121(11–12): 1264–1272. DOI: 10.1289/ehp.1306556.

Moyo C, Kaasbøll J, Nielsen P, et al. (2016) The Information Transparency Effects of Introducing League Tables in the Health System in Malawi. *The Electronic Journal of Information Systems in Developing Countries* 75(1): 1–16. DOI: 10.1002/j.1681-4835.2016.tb00544.x.

Murhekar MV, Kamaraj P, Kumar MS, et al. (2019) Burden of dengue infection in India, 2017: a cross-sectional population based serosurvey. *The Lancet Global Health* 7(8): e1065–e1073. DOI: 10.1016/S2214-109X(19)30250-5.

Mutale W, Chintu N, Amoroso C, et al. (2013) Improving health information systems for decision making across five sub-Saharan African countries: Implementation strategies from the African Health Initiative. *BMC Health Services Research* 13(S2): S9. DOI: 10.1186/1472-6963-13-S2-S9.

Mutheneni SR, Morse AP, Caminade C, et al. (2017) Dengue burden in India: recent trends and importance of climatic parameters. *Emerging Microbes & Infections* 6(1): 1–10. DOI: 10.1038/emi.2017.57.

National Center for Vector Borne Diseases Control (2022) DENGUE/DHF SITUATION IN INDIA. Available at: https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=431&lid=3715 (accessed 2 December 2022).

National Centre for Disease Control, Directorate General of Health Services (2022) Integrated Disease Surveillance Programme(IDSP). Available at: https://idsp.nic.in/index4.php?lang=1&level=0&linkid=313&lid=1592 (accessed 11 November 2022).

National Data Sharing and Accessibility Policy | Department Of Science & Technology (2022). Available at: https://dst.gov.in/national-data-sharing-and-accessibility-policy-0 (accessed 30 July 2022).

NCBI (2021) Data Science. Available at: https://www.ncbi.nlm.nih.gov/mesh/2028050 (accessed 30 January 2021).

Nori-Sarma A, Gurung A, Azhar G, et al. (2017) Opportunities and Challenges in Public Health Data Collection in Southern Asia: Examples from Western India and Kathmandu Valley, Nepal. *Sustainability* 9(7): 1106. DOI: 10.3390/su9071106.

Nutley T and Reynolds HeidiW (2013) Improving the use of health data for health system strengthening. *Global Health Action* 6(1): 20001. DOI: 10.3402/gha.v6i0.20001.

Nutley T, Gnassou L, Traore M, et al. (2014) Moving data off the shelf and into action: an intervention to improve data-informed decision making in Côte d'Ivoire. *Global Health Action* 7(1): 25035. DOI: 10.3402/gha.v7.25035.

NVBDCP (2021) DENGUE/DHF SITUATION IN INDIA. Available at: https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=431&lid=3715 (accessed 30 January 2021).

Nwankwo B and Sambo MN (2018) Can training of health care workers improve data management practice in health management information systems: a case study of primary health care facilities in Kaduna State, Nigeria. *The Pan African Medical Journal* 30: 289. DOI: 10.11604/pamj.2018.30.289.15802.

Ong J, Liu X, Rajarethinam J, et al. (2018) Mapping dengue risk in Singapore using Random Forest. *PLoS neglected tropical diseases* 12(6): e0006587. DOI: 10.1371/journal.pntd.0006587.

Ooi E-E and Gubler DJ (2009) Global spread of epidemic dengue: the influence of environmental change. *Future Virology* 4(6). 6. Future Medicine: 571–580. DOI: 10.2217/fvl.09.55.

Open Government Data (OGD) Platform India (2022) Open Government Data (OGD) Platform India. Available at: https://data.gov.in (accessed 29 September 2022).

Oracle India (2021) What Is Big Data? Available at: https://www.oracle.com/in/big-data/what-is-big-data/ (accessed 6 April 2021).

O'Sullivan D and Unwin D (2014) *Geographic Information Analysis*. Wiley. Available at: https://books.google.co.in/books?id=yVUzBAAAQBAJ.

P M Ashburn and Charles F Craig (2004) Experimental Investigations Regarding the Etiology of Dengue. *The Journal of Infectious Diseases* 189(9): 1744–1783. DOI: 10.1086/383418.

Pei S, Kandula S, Yang W, et al. (2018) Forecasting the spatial transmission of influenza in the United States. *Proceedings of the National Academy of Sciences of the United States of America* 115(11): 2752–2757. DOI: 10.1073/pnas.1708856115.

Pfeiffer DU, Robinson TP, Stevenson M, et al. (2008) *Spatial Analysis in Epidemiology*. OUP Oxford. Available at: https://books.google.co.in/books?id=wQlREAAAQBAJ.

Phan HTT, Borca F, Cable D, et al. (2020) Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort. *Scientific Reports* 10(1): 10164. DOI: 10.1038/s41598-020-66925-7.

Phanitchat T, Zhao B, Haque U, et al. (2019) Spatial and temporal patterns of dengue incidence in northeastern Thailand 2006–2016. *BMC Infectious Diseases* 19(1): 743. DOI: 10.1186/s12879-019-4379-3.

Ramadona AL, Tozan Y, Lazuardi L, et al. (2019) A combination of incidence data and mobility proxies from social media predicts the intra-urban spread of dengue in Yogyakarta, Indonesia. *PLOS Neglected Tropical Diseases* Werneck GL (ed.) 13(4): e0007298. DOI: 10.1371/journal.pntd.0007298.

Randall SM, Ferrante AM, Boyd JH, et al. (2013) The effect of data cleaning on record linkage quality. *BMC Medical Informatics and Decision Making* 13(1): 64. DOI: 10.1186/1472-6947-13-64.

Ren H, Wu W, Li T, et al. (2019) Urban villages as transfer stations for dengue fever epidemic: A case study in the Guangzhou, China. *PLOS Neglected Tropical*

Diseases Reiner RC (ed.) 13(4): e0007350. DOI: 10.1371/journal.pntd.0007350.

Rueda LM, Patel KJ, Axtell RC, et al. (1990) Temperature-Dependent Development and Survival Rates of Culex quinquefasciatus and Aedes aegypti (Diptera: Culicidae). *Journal of Medical Entomology* 27(5): 892–898. DOI: 10.1093/jmedent/27.5.892.

Sánchez-González G, Condé R, Noguez Moreno R, et al. (2018) Prediction of dengue outbreaks in Mexico based on entomological, meteorological and demographic data. *PloS One* 13(8): e0196047. DOI: 10.1371/journal.pone.0196047.

Sánchez-Hernández D, Aguirre-Salado CA, Sánchez-Díaz G, et al. (2019) Modeling spatial pattern of dengue in North Central Mexico using survey data and logistic regression. *International Journal of Environmental Health Research*: 1–17. DOI: 10.1080/09603123.2019.1700938.

Shil P (2019) Rainfall and dengue occurrences in India during 2010–2016. *Biomedical Research Journal* 6(2): 56. DOI: 10.4103/BMRJ.BMRJ_15_19.

Singh G, Tilak R and Kaushik S (2019) Bio-eco-social determinants of Aedes breeding in field practice area of a medical college in Pune, Maharashtra. *Indian Journal of Public Health* 63(4): 324. DOI: 10.4103/ijph.IJPH_296_18.

Singh G, Mitra A and Soman B (2022) Development and use of a reproducible framework for spatiotemporal climatic risk assessment and its association with decadal trend of dengue in India. *Indian Journal of Community Medicine* 47(1): 50. DOI: 10.4103/ijcm.ijcm_862_21.

Siriyasatien P, Chadsuthi S, Jampachaisri K, et al. (2018) Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes. *IEEE Access* 6: 53757–53795. DOI: 10.1109/ACCESS.2018.2871241.

Smeets HM, de Wit NJ and Hoes AW (2011) Routine health insurance data for scientific research: potential and limitations of the Agis Health Database. *Journal of Clinical Epidemiology* 64(4): 424–430. DOI: 10.1016/j.jclinepi.2010.04.023.

Stolerman LM, Maia PD and Kutz JN (2019) Forecasting dengue fever in Brazil: An assessment of climate conditions. *PLOS ONE* Samy AM (ed.) 14(8): e0220106. DOI: 10.1371/journal.pone.0220106.

Strimbu BM, Amarioarei A and Paun M (2017) A parsimonious approach for modeling uncertainty within complex nonlinear relationships. *Ecosphere* 8(9). DOI: 10.1002/ecs2.1945.

Sun G, Nakayama Y, Dagdanpurev S, et al. (2017) Remote sensing of multiple vital signs using a CMOS camera-equipped infrared thermography system and its clinical application in rapidly screening patients with suspected infectious diseases. *INTERNATIONAL JOURNAL OF INFECTIOUS DISEASES* 55: 113–117. DOI: 10.1016/j.ijid.2017.01.007.

Swain S, Bhatt M, Pati S, et al. (2019) Distribution of and associated factors for dengue burden in the state of Odisha, India during 2010–2016. *Infectious Diseases of Poverty* 8(1): 31. DOI: 10.1186/s40249-019-0541-9.

Telle O, Vaguet A, Yadav NK, et al. (2016) The Spread of Dengue in an Endemic Urban Milieu–The Case of Delhi, India. *PLOS ONE* Costa C (ed.) 11(1): e0146539. DOI: 10.1371/journal.pone.0146539.

Titus Muurlink O, Stephenson P, Islam MZ, et al. (2018) Long-term predictors of dengue outbreaks in Bangladesh: A data mining approach. *Infectious Disease Modelling* 3: 322–330. DOI: 10.1016/j.idm.2018.11.004.

Tukey JW (1977) *Exploratory Data Analysis*. Addison-Wesley series in behavioral science. Reading, Mass: Addison-Wesley Pub. Co.

Tun-Lin W, Burkot TR and Kay BH (2000) Effects of temperature and larval diet on development rates and survival of the dengue vector Aedes aegypti in north Queensland, Australia. *Medical and Veterinary Entomology* 14(1): 31–37. DOI: 10.1046/j.1365-2915.2000.00207.x.

Van den Broeck J, Argeseanu Cunningham S, Eeckels R, et al. (2005) Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. *PLoS Medicine* 2(10): e267. DOI: 10.1371/journal.pmed.0020267.

Van der Aalst (2016) *Process Mining: Data Science in Action*. 2nd edition. New York, NY: Springer Berlin Heidelberg.

van der Aalst W (2016) Data Science in Action. In: van der Aalst W (ed.) *Process Mining: Data Science in Action*. Berlin, Heidelberg: Springer, pp. 3–23. DOI: 10.1007/978-3-662-49851-4_1.

Verma M, Kishore K, Kumar M, et al. (2018) Google Search Trends Predicting Disease Outbreaks: An Analysis from India. *Healthcare Informatics Research* 24(4): 300. DOI: 10.4258/hir.2018.24.4.300.

Vincenti-Gonzalez MF, Tami A, Lizarazo EF, et al. (2018) ENSO-driven climate variability promotes periodic major outbreaks of dengue in Venezuela. *Scientific Reports* 8(1): 5727. DOI: 10.1038/s41598-018-24003-z.

Vissoci JRN, Rocha TAH, Silva NC da, et al. (2018) Zika virus infection and microcephaly: Evidence regarding geospatial associations. *PLoS neglected tropical diseases* 12(4): e0006392. DOI: 10.1371/journal.pntd.0006392.

Volkova S, Ayton E, Porterfield K, et al. (2017) Forecasting influenza-like illness dynamics for military populations using neural networks and social media. *PloS One* 12(12): e0188941. DOI: 10.1371/journal.pone.0188941.

Wagenaar BH, Gimbel S, Hoek R, et al. (2015) Effects of a health information system data quality intervention on concordance in Mozambique: time-series analyses from 2009–2012. *Population Health Metrics* 13(1): 9. DOI: 10.1186/s12963-015-0043-3.

Wagenaar BH, Sherr K, Fernandes Q, et al. (2016) Using routine health information systems for well-designed health evaluations in low- and middle-income countries. *Health Policy and Planning* 31(1): 129–135. DOI: 10.1093/heapol/czv029.

Wang X, Tang S, Wu J, et al. (2019) A combination of climatic conditions determines major within-season dengue outbreaks in Guangdong Province, China. *Parasites & Vectors* 12(1): 45. DOI: 10.1186/s13071-019-3295-0.

Wangdi K, Clements ACA, Du T, et al. (2018) Spatial and temporal patterns of dengue infections in Timor-Leste, 2005–2013. *Parasites & Vectors* 11(1): 9. DOI: 10.1186/s13071-017-2588-4.

Whiteman A, Desjardins MR, Eskildsen GA, et al. (2019) Detecting space-time clusters of dengue fever in Panama after adjusting for vector surveillance data. *PLOS Neglected Tropical Diseases* Scarpino SV (ed.) 13(9): e0007266. DOI: 10.1371/journal.pntd.0007266.

Wilhelm JA, Qiu M, Paina L, et al. (2019) The impact of PEPFAR transition on HIV service delivery at health facilities in Uganda. *PLOS ONE* Rockers P (ed.) 14(10): e0223426. DOI: 10.1371/journal.pone.0223426.

Withanage GP, Viswakula SD, Nilmini Silva Gunawardena YI, et al. (2018) A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka. *Parasites & Vectors* 11(1): 262. DOI: 10.1186/s13071-018-2828-2.

World Health Organization (2007) *Everybody's Business: Strengthening Health Systems to Improve Health Outcomes : WHO's Framework for Action.* Geneva: World Health Organization.

World Health Organization (2011) *Comprehensive Guidelines for Prevention and Control of Dengue and Dengue Haemorrhagic Fever*. Rev. and expanded. ed. SEARO Technical publication series no. 60. New Delhi, India: World Health Organization Regional Office for South-East Asia.

World Health Organization (2012) Global Strategy for Dengue prevention and control.

World Health Organization (2022a) Dengue and severe dengue. Available at: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue (accessed 24 February 2023).

World Health Organization (2022b) What is dengue? World Health Organization. Available at: http://www.who.int/denguecontrol/disease/en/ (accessed 8 May 2022).

Xu J, Xu K, Li Z, et al. (2020) Forecast of Dengue Cases in 20 Chinese Cities Based on the Deep Learning Method. *International Journal of Environmental Research and Public Health* 17(2): 453. DOI: 10.3390/ijerph17020453.

Xu Z, Bambrick H, Yakob L, et al. (2020) High relative humidity might trigger the occurrence of the second seasonal peak of dengue in the Philippines. *Science*

of The Total Environment 708: 134849. DOI: 10.1016/j.scitotenv.2019.134849.

Yan Z and Haimes YY (2011) Risk-based multiobjective resource allocation in hierarchical systems with multiple decisionmakers. Part I: Theory and methodology. *Systems Engineering* 14(1): 1–16. DOI: 10.1002/sys.20159.

Yang W, Wen L, Li S-L, et al. (2017) Geospatial characteristics of measles transmission in China during 2005-2014. *PLoS computational biology* 13(4): e1005474. DOI: 10.1371/journal.pcbi.1005474.

Yuan M, Boston-Fisher N, Luo Y, et al. (2019) A systematic review of aberration detection algorithms used in public health surveillance. *Journal of Biomedical Informatics* 94: 103181. DOI: 10.1016/j.jbi.2019.103181.

Zahirul Islam M, Rutherford S, Phung D, et al. (2018) Correlates of Climate Variability and Dengue Fever in Two Metropolitan Cities in Bangladesh. *Cureus*. DOI: 10.7759/cureus.3398.

Zambrana JV, Bustos Carrillo F, Burger-Calderon R, et al. (2018) Seroprevalence, risk factor, and spatial analyses of Zika virus infection after the 2016 epidemic in Managua, Nicaragua. *Proceedings of the National Academy of Sciences of the United States of America* 115(37): 9294–9299. DOI: 10.1073/pnas.1804672115.

Zhang J and Nawata K (2018) Multi-step prediction for influenza outbreak by an adjusted long short-term memory. *Epidemiology and Infection* 146(7): 809–816. DOI: 10.1017/S0950268818000705.

Zhang Q, Gao J, Wu JT, et al. (2022) Data science approaches to confronting the COVID-19 pandemic: a narrative review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 380(2214): 20210127. DOI: 10.1098/rsta.2021.0127.

Zhang Qin, Chen Y, Fu Y, et al. (2019) Epidemiology of dengue and the effect of seasonal climate variation on its dynamics: a spatio-temporal descriptive analysis in the Chao-Shan area on China's southeastern coast. *BMJ Open* 9(5): e024197. DOI: 10.1136/bmjopen-2018-024197.

Zhang Y, Wang T, Liu K, et al. (2016) Developing a Time Series Predictive Model for Dengue in Zhongshan, China Based on Weather and Guangzhou Dengue Surveillance Data. *PLOS Neglected Tropical Diseases* Scarpino SV (ed.) 10(2). 2: e0004473. DOI: 10.1371/journal.pntd.0004473.

Zheng L, Ren H-Y, Shi R-H, et al. (2019) Spatiotemporal characteristics and primary influencing factors of typical dengue fever epidemics in China. *Infectious Diseases of Poverty* 8(1). 1: 24. DOI: 10.1186/s40249-019-0533-9.

Zhu B, Liu J, Fu Y, et al. (2018) Spatio-Temporal Epidemiology of Viral Hepatitis in China (2003-2015): Implications for Prevention and Control Policies. *International Journal of Environmental Research and Public Health* 15(4). DOI: 10.3390/ijerph15040661.

Zhu G, Xiao J, Liu T, et al. (2019) Spatiotemporal analysis of the dengue outbreak in Guangdong Province, China. *BMC Infectious Diseases* 19(1): 493. DOI: 10.1186/s12879-019-4015-2.

Zodpey S and Negandhi H (2016) Improving the quality and use of routine health data for decision-making. *Indian Journal of Public Health* 60(1): 1. DOI: 10.4103/0019-557X.177248.

ANNEXURES

LIST OF PUBLICATIONS FROM THESIS

Singh G and Soman B (2021) Spatiotemporal Epidemiology and Forecasting of Dengue in the state of Punjab, India: Study Protocol.,. Spatial and Spatio-temporal Epidemiology: 100444. DOI: 10.1016/j.sste.2021.100444.

Singh G, Soman B and Mitra A (2021) A Systematic Approach to Cleaning Routine Health Surveillance Datasets: An Illustration Using National Vector Borne Disease Control Programme Data of Punjab, India. arXiv:2108.09963 [cs]. Available at: http://arxiv.org/abs/2108.09963 (accessed 24 September 2021).

Singh G, Mitra A and Soman B (2022) Development and use of a reproducible framework for spatiotemporal climatic risk assessment and its association with decadal trend of dengue in India. Indian Journal of Community Medicine 47(1): 50. DOI: 10.4103/ijcm.ijcm_862_21.

Singh G, Soman B and Grover GS (2022) Development and use of open-source algorithms for space-time emerging hotspot analysis of routine dengue NVBDCP data in Punjab, India. International Journal Of Community Medicine And Public Health 10(1): 148. DOI: 10.18203/2394-6040.ijcmph20223288.

Singh G, Soman B and Grover GS (2023) Exploratory Spatio-Temporal Data Analysis (ESTDA) of Dengue and its association with climatic, environmental, and sociodemographic factors in Punjab, India. Ecological Informatics 75: 102020. DOI: 10.1016/j.ecoinf.2023.102020.

CURRICULUM VITAE

Format for CV of the Investigators

Last Name SINGH	**First Name** GURPREET	**Middle Name**
Date of Birth (dd/mm/yy) 08/01/85		**Sex** MALE
Study Site Affiliation (e.g. Principal Investigator, Co-Investigator, Coordinator) PhD Student		
Professional Mailing Address **(Include Institution name)**		**Study Site Address** **(Include Institution name)**
Achuta Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, Kerala - 695011		Achuta Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, Kerala - 695011
Telephone (Office):		**Mobile Number:** 8552055667
Telephone (Residence):		**Email** drgurpreet_amc@sctimst.ac.in

Academic Qualifications (Most recent qualification first)

Degree/Certificate	**Year**	**Institution, Country**
DNB Preventive and Social Medicine	2016	National Board of Examinations
MD Community Medicine	2015	Maharashtra University of Health Sciences
MBBS	2006	Maharashtra University of Health Sciences

Details of professional registration : (MCI/State Registration/Bar Council/DCI/etc including Registration Number and Year of Registration MCI- 08/32461, 2008

Current and previous positions (most recent position first)

Month and Year	**Title**	**Institution/Company, Country**
Jan 2019-till date	PhD Scholar	AMCHSS, SCTIMST, Kerala, India
Jun 2016-Jan 2019	Officer Commanding	Station Health Organization, Jalandhar Cantt, Punjab, India
Jun 2015-Jun 2016	Assistant Professor	Armed Forces Medical College, India
May 2012-May 2015	Resident	Armed Forces Medical College, India
Feb 2007-May 2012	Medical Officer	Armed Forces Medical Services, India

Brief summary of relevant research experience:
Total publications in peer reviewed indexed journals: 46.
Total R codes published: 20

Signature:	Date: Place: Trivandrum

APPENDICES

APPENDIX A – ETHICS COMMITTEE APPROVAL

श्री चित्रा तिरुनाल आयुर्विज्ञान और प्रौद्योगिकी संस्थान, त्रिवेंद्रम - 695 011, केरल, भारत
SREE CHITRA TIRUNAL INSTITUTE FOR MEDICAL SCIENCES AND TECHNOLOGY
TRIVANDRUM - 695 011, KERALA, INDIA
(एक राष्ट्रीय महत्व का संस्थान, विज्ञान एवं प्रौद्योगिकी विभाग, भारत सरकार)
(An Institution of National Importance, Department of Science and Technology, Government of India)
टेलीफॉन नं./Telephone No.: 0471-2443152 फैक्स/Fax: 0471-2446433, 2550728
ई-मेल/E-mail: sct@sctimst.ac.in वेबसाइट/Website: www.sctimst.ac.in

Institutional Ethics Committee

(IEC Regn No. ECR/189/Inst/KL/2013/RR-16)

SCT/IEC/IEC-1653/DECEMBER-2020 19.12.2020

Dr Gurpreet Singh
Ph.D. Scholar AMCHSS, SCTIMST, Trivandrum

Dear Dr Gurpreet Singh,
Thank you for submitting documents related to your proposal titled ""**Spatiotemporal epidemiology and forecasting of dengue in the state of Punjab, India**". (IEC/IEC-1653)" to the IEC for review.

The following documents were reviewed:

1. Check list
2. Covering letter addressed to Chairman dated 09.10.2020 by the Guide
3. IEC application
4. TAC Approval date 30.09.2020
5. Research Proposal
6. Letter from Directorate of Health & Family Welfare, Chandigrah reg. permission for the use of data related dengue, dated15/01/2020
7. CV of Investigator Dr.Gurpreet singh with MCI number
8. CV of Dr.Biju Soman with TCMC number

IEC Recommendations

The study is recommended for approval.

The following members of the Institutional Ethics Committee participated in the discussions held virtually on Dec 18 2020, at the offices and residences of the members

SL. No.	Member Name	Highest Degree	Gender	Scientific /Non Scientific	Affiliation with Institution(s)
1.	Dr. R V G Menon	M Tech, PhD	Male	Lay Person (Chairman)	No
2.	Dr. Rema M. N	MD	Female	Basic Medical Scientist	No
3.	Dr. Kala Kesavan. P	MBBS, MD	Female	Basic Medical Scientist	No
4.	Dr. K R S Krishnan	M.E., Ph.D.	Male	Medical Technology	No
5.	Dr. Harikrishna Varma PR	Ph.D(Materials Science)	Male	Medical Technology	Yes
6.	Dr. S S Giri Sankar	LL.M. Ph.D.	Male	Legal Expert	No
7.	Dr. Anand Kumar A	MD, DM	Male	Clinician	No
8.	Dr. Aneesh V Pillai	BA. LLB (Hons.), LLM, Ph. D, SET (Law)	Male	Legal Expert	No
9.	Smt. Sathi Nair	MA (English Literature)	Female	Lay Person	No
10.	Dr. P. Manickam	BSMS, MSc (Epid).,PhD	Male	Health Science Expert/ Social Scientist	No
11	Dr.Raman Kutty V	MD (Padiatrics), Mphil, MPH	Male	Health Science Expert/Clinician	No
12.	Dr. Harikrishnan S	MD, DM (Cardiology) DNB (Cardiology)	Male	Clinician	Yes
13.	Dr. Christina George	MD Psychiatry	Female	Clinician	No
14.	Mr. Satheesh Chandran	MSW, PGDPM	Male	Lay person/ NGO/ Social Scientist	No
15.	Dr. Mala Ramanathan	PhD	Female	Social Scientist (Member Secretary)	Yes

IEC Decision
The IEC approved the conduct of the study in the present form.

Remarks:
The Institutional Ethics Committee expects to be informed about the progress of the study, any SAE occurring in the course of the study, any changes in the protocol and patient information/informed consent and asks to be provided a copy of the final report.
There was no member of the study team who participated in voting / decision making process. The ethics committee is organized and operated according to the requirements of Good Clinical Practice and the requirements of the Indian Council of Medical Research (ICMR).
Sincerely,

Mala Ramanathan
Member Secretary, IEC

APPENDIX D - PUBLICATIONS

Spatial and Spatio-temporal Epidemiology 39 (2021) 100444

Contents lists available at ScienceDirect

Spatial and Spatio-temporal Epidemiology

journal homepage: www.elsevier.com/locate/sste

Spatiotemporal epidemiology and forecasting of dengue in the state of Punjab, India: Study protocol

Gurpreet Singh, Biju Soman *

[a] Achutha Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, India.

ARTICLE INFO

Keywords:
Dengue
Spatiotemporal
Machine learning
Data science
Routine data
Surveillance

ABSTRACT

Dengue burden in India is a major public health problem. The present study has been designed to understand mechanisms by which routine data generate evidence. Secondary data analysis of routine datasets to understand spatiotemporal epidemiology and forecast dengue will be conducted. Data science approach will be adopted to generate a reproducible framework in the R environment. The lab-confirmed dengue reported by the state health authorities from 01 January 2015 to 31 December 2019 will be included. Multiple climatic variables from satellite imagery, climatic models, vegetation and built-up indices, and sociodemographic variables will be explored as risk factors. Exploratory data analysis followed by statistical analysis and machine learning will be performed. Data analysis will include geospatial information analysis, time series analysis, and spatiotemporal analysis. The study will provide value addition to the existing disease surveillance mechanisms by developing a framework for incorporating multiple routine data sources available in the country.

1. Introduction

Among infectious diseases, more than 17% are attributed to vector-borne diseases (Vector-borne diseases, 2020). Dengue is the most prevalent viral infection transmitted by *Aedes* mosquitoes and globally, its incidence has increased more than 15 times in the past two decades. The risk of acquiring dengue infection is present in more than 129 countries, with an estimated 96 million clinical cases (Vector-borne diseases, 2020; World Health Organization, 2020). A recent multicentric study to estimate dengue sero-surveillance in India found heterogeneous transmission of dengue infection in the country with high disease burden in north, south, and west parts of the country (Murhekar et al., 2019). According to the National Vector Borne Disease Control Program (NVBDCP), the state of Punjab in India reported the highest number of dengue cases among all states in 2015 and has been reporting approximately 10,000 dengue cases every year since then (Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India, 2021a).

The transmission of the dengue includes humans and mosquitoes. Humans play a vital role as a source of infection and mosquitoes act as vectors for disease transmission in the community (World Health Organization, 2020). Climatic and Environmental factors have established a role in dengue transmission dynamics resulting from their effect on the vector bionomics, viral development, and vector host interactions (Farrar and Manson, 2014). Temperature, the most prominent climatic factor associated with dengue incidence, directly affects the availability of *Aedes* water habitat sources, developmental process, survivability in the environment, reproduction rates, and dengue virus replication and transmission rates (Morin et al., 2013). Precipitation, independently and in interaction with temperature determines evaporation rates, thus habitat availability for immature stages of the mosquito life cycle. Land use, land cover, and vegetation influence micro-climatic conditions affecting mosquito density (J. Xu et al., 2020a; Xu et al., 2020b; Zhang et al., 2016; Zheng et al., 2019). The conduciveness of the environment for the transmission of dengue is also dependent on the sociodemographic profile of the community (Singh et al., 2019). Poor public health infrastructure, lack of civic amenities, rapid urbanization, and other sociodemographic factors affect the availability of breeding places for mosquitoes and impact vector-human interactions (Banu et al., 2011; Ooi and Gubler, 2009).

Public health surveillance aims at providing information at a timely interval for action (Last and International Epidemiological Association, 2001). National Vector Borne Disease Control Programme (NVBDCP), and Integrated Disease Surveillance Programme, India are the nodal

Source(s) of support in the form of grants, equipment, drugs, or all of these: NILSubmissions and previous reports: NIL
* Corresponding author.
E-mail address: bijusoman@sctimst.ac.in (B. Soman).

https://doi.org/10.1016/j.sste.2021.100444
Received 22 February 2021; Received in revised form 2 July 2021; Accepted 21 July 2021
Available online 24 July 2021

programs for the surveillance and control of vector-borne diseases in the country (Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India, 2021b). The early detection and response mechanisms in the country are based on the reported disease burden as well as inputs from media and other informal sources such as rumor registries (National center for Disease Control, Directorate General of Health Services, 2021). Expansion on Geographical Information Systems (GIS), signal detection algorithms, and linking to databases from other non-health sectors have been recommended for strengthening disease surveillance in the country (Directorate General of Health Services, India, 2015).

In the recent past, data science has developed as a novel discipline (van der Aalst, 2016). The term data science has been defined as "an interdisciplinary field involving processes, theories, concepts, tools, and technologies, that enable the review, analysis, and extraction of valuable knowledge and information from structured and unstructured (raw) data (NCBI, 2021). Understanding of disease epidemiology, application of data science, and resulting advances in informatics have enabled researchers globally to develop aberration detection and forecasting algorithms for strengthening disease surveillance systems (Yuan et al., 2019). With the ever-increasing generation of data and the use of technologies in understanding the epidemiology of diseases, the scope and application of data analytics in healthcare have increased manifold. Tools such as geospatial information analysis, time series analysis, and machine learning algorithms have been used for understanding disease patterns as well as for forecasting disease outbreaks (Bouzillé et al., 2018; Chae et al., 2018; Ong et al., 2018; Pei et al., 2018; Sánchez-González et al., 2018; Volkova et al., 2017; Wang et al., 2019; Withanage et al., 2018; Zhang and Nawata, 2018). Further, the application of data science tools and technologies have been demonstrated to map the risk of diseases across populations with a view of enabling efficient utilization of constrained public health resources (Minale and Alemu, 2018; Zambrana et al., 2018). However, the application of these tools and technologies in understanding the spatiotemporal epidemiology of dengue has been limited in the Indian sub-continent. Thus, the present study has been designed to understand the mechanisms by which routine data generate evidence for public health decision making. The objectives of the study are to describe the spatiotemporal distribution of dengue and its selected risk factors, to explore associations between dengue and climatic, environmental, and sociodemographic risk factors, to develop a dengue forecasting model using routine data sources, and thus, based on the learnings, develop a reproducible framework for routine data-based vector-borne disease forecasting models. Such models have the potential for the development of evidence-informed public health decision support mechanisms in low- and middle-income countries, and thus, strengthen surveillance by complementing existing mechanisms.

2. Material and methods

2.1. Study settings

The study will be conducted in the state of Punjab, India. Punjab is a northern state of India with a total area of 50,362 square kilometers and a total population of 2,77,43,338. It extends from the latitudes 29.30° - 32.32° north and longitudes 73.55° - 76.50° east. The state is divided into administrative units known as districts. Further, each district is divided into health blocks. According to Census 2011, there were 20 districts and 142 blocks in the state. Additional two districts were formed on 27 July 2011, one each from *Firozpur* and *Gurdaspur* districts respectively. The state experiences extreme climatic seasonal patterns viz. winters, rains, and summers, and has an average elevation (range) of 300 m (180 to >500 m) above sea level (Government of Punjab, India, 2021).

2.2. Study design

The proposed study includes secondary data analysis of routine datasets incorporating a data science approach. Spatiotemporal analysis of data collected from multiple routine data sources in the health sector and non-healthcare sectors (population enumeration data, remote sensing data, and data from global climatic models) will be carried out. These historic datasets will be used to train computer programs in developing a Bayesian model for forecasting dengue. The knowledge gained and learnings will be utilized for the development of a framework for the routine data-based Vector-Borne Disease (VBD) forecasting models.

2.3. Study population

The lab-confirmed dengue cases as reported by the state health authorities from 01 January 2015 to 31 December 2019 will be analyzed. Those cases reported from outside the state will be excluded. Details of lab-confirmed dengue cases reported by the state are represented in Table 1.

2.4. Plan for data collection and analysis

The plan for data collection and analysis is represented schematically in Fig. 1. Data collected from routine data sources will be extracted and pre-processed to transform datasets into analyzable tidy formats. Exploratory data analysis followed by statistical analysis and application of machine learning algorithms will be carried out to understand the epidemiology of dengue in the state and develop a dengue forecasting model using a reproducible framework.

2.4.1. Data collection.

Data sources for the proposed study will include routine dengue surveillance data (line listing of lab-confirmed dengue cases) from National Vector Borne Disease Control Programme, Directorate of Health Services, Punjab; multiple climatic variables from remote sensing data (research level) provided by Modern-Era Retrospective analysis for Research and Applications (MERRA-2), Integrated Multi-satellitE Retrievals for GPM (IMERG), and Moderate Resolution Imaging Spectroradiometer (MODIS) datasets; Vegetation indices and elevation data provided by National Remote Sensing center on Bhuvan web portal; normalized difference built-up index from Landsat 8 Operational Land Imager (OLI) imagery data; Sociodemographic variables from Census 2011, NFHS-4, and Socioeconomic Data and Applications Center (SEDAC) datasets; population projections (National Commission on Population, 2019); and spatial datasets provided by Punjab Remote Sensing Authority and geospatial resources at Achutha Menon center for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology. The health records in the line list will be checked for duplicate entries prior to anonymization during the data collection. Duplicates will be looked into for common name, age, sex, mobile number, address, and date of testing. The data will be anonymized using 'scrypt' algorithm available from *epitrix* library.

2.4.2. Data extraction

The date of testing, admission, and discharge will be extracted along

Table 1
Study population.

S No	Year	Line listing of cases (NVBDCP, Punjab)
1	2015	15,543
2	2016	9893
3	2017	15,406
4	2018	15,569
5	2019	10,170
	Total	66,581

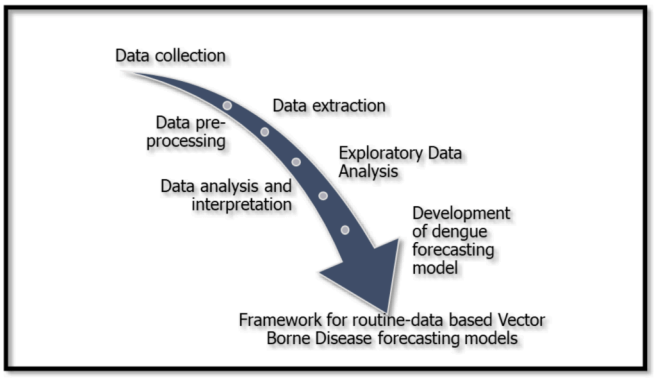


Fig. 1. Data collection and analysis plan.

with details on age, sex, type of test performed, hospitalizations, deaths, and location from dengue line list data. The raw addresses in the line list will be geocoded using google API after appending district and sub-district details. The variable wise data extraction template for environmental and sociodemographic datasets, including satellite imagery data extraction are presented in Table 2.

2.4.3. Data pre-processing

Tidy datasets for analysis will be prepared using reproducible data pre-processing algorithms. Data cleaning to develop analysis ready formats for machine learning, and batch pre-processing of the satellite imagery data for preparation of spatial and time series attribute data will be carried out. Data transformations for individual variables (centering, scaling, and skewness transformations), data reduction, and feature extraction and selection among other methods for pre-processing routine data will be carried out, as required.

2.4.4. Exploratory data analysis

Exploratory data analysis will include visualization of the datasets to understand the data structure. It will be based on spatial plots, time series plots, and spatiotemporal visualizations. Also, summary measures for spatial variation, clustering, time series components, and spatio-temporal patterns of dengue incidence, environmental, climatic, and sociodemographic factors will be explored.

2.4.5. Data analysis and interpretation

Data analysis will include geospatial information analysis, time series analysis, spatiotemporal analysis, and the use of machine learning algorithms for developing dengue forecasting models. Spatial distribution, estimation of spatial clustering at global and local levels, estimation of spatial variation in risk, and spatial modeling will be carried out based on both frequentist and Bayesian approaches. Dengue clusters in the state will be defined as regions where the dengue incidence is significantly higher than the rest of the state. Global and local Moran's I statistic and Getis Ord Gi* statistic will be calculated for estimates of clustering, outlier analysis, and hotspot analysis, respectively. Neighborhoods in space and time will be defined on the basis of contiguity/ fixed distance bands and on specified timestamps, based on findings of exploratory analysis. Emerging hotspot analysis will be carried out to determine spatiotemporal characteristics. The time series analysis component in the study will include time-series decomposition analysis for the development of the state-space model. The study will include estimates of annual, quarterly/ seasonal, monthly, fortnightly, and weekly intervals. The autocorrelation characteristics will be explored for the development of Auto Regression and Moving Average based model. For the machine learning component, the forecasting model will be developed on the training dataset and validated on the testing dataset. Being count data, Poisson distribution will be assumed. Further modifications based on expected and observed distribution patterns (Zero Inflated Poisson model, negative binomial models, Integrated Nested Laplace Approximation (INLA) based models, etc.) will be explored. The best fit model shall be based on model evaluation parameters such as Mean Squared Errors (MSE), Akaike Information Criteria (AIC), and adjusted R^2 as indicated. The residual analysis will be performed to validate model assumptions.

2.5. Ethical considerations

The proposed study will be carried out following ICMR guidelines. The study has been approved by the Technical Advisory Committee (AMCHSS/2020/5/001) and Institutional Ethics Committee (IEC/IEC-1653; IEC Reg. No. ECR/189/Inst/KL/2013/RR-16) dated 19/12/2020. Permission from the Directorate of Health Services, Punjab; Punjab Remote Sensing Authority has been obtained. The study has also been registered at the Clinical Trials Registry of India (CTRI/2021/01/030,245). Permission for the use of block-level administrative boundaries spatial dataset has been obtained from Punjab Remote Sensing Authority and necessary registrations for bulk downloads and access to remote sensing data from Bhuvan web portal of National Remote Sensing center, India, and earth data, NASA has been carried out. The privacy and confidentiality of the datasets will be maintained, and geo-masking of geocoded data will be carried out during the dissemination of findings.

2.6. Data storage transfer and management

All the datasets will be password encrypted, and data access will be provided to only the Principal Investigator and Supervisor. Weekly back-

Table 2
Data source and extraction template for exposure variables.

S No	Variable	Source	File format	Spatial resolution	Temporal resolution
1	Multiple climatic data*	MERRA-2	Excel	0.5°	Daily
2	Precipitation	GPM IMERG Final Precipitation L3	NetCDF	0.1°	Daily
3	Temperature	MOD11C2 Version 6 (LST&E)	HDF	0.05°	8 days
4	Projected Population	Calculated**	Excel	Sub-district	Annual
5	Sociodemographic variables***	Primary Census Abstract	Excel	Sub-district	–
6	Built index	Landsat8	Geo-Tiff	30m	Annual
7	Elevation	Indian National DEM (Version-2)/ CartoDEM	Geo-Tiff	10m	–
8	Land Vegetation	OCM2: Filter NDVI Product	Geo-Tiff	1080m	15 days
9	Spatial Datasets	PRSC, Ludhiana	Shapefile	Sub-district	–

* 10-meter air temperature, 2-meter air temperature, 10-meter specific humidity, 2-meter specific humidity, Wet-bulb temperature at 2 m, Total precipitable water vapor, Air temperature at 250 hPa, Air temperature at 500 hPa, Air temperature at 850 hPa, Dew point temperature at 2 m, Specific humidity at 250 hPa, Specific humidity at 500 hPa, Specific humidity at 850 hPa.

** Population projections will be calculated based on annualized growth rates and population estimates available from Census 2011, and adjusted to population projections for the respective years during the study period (National Commission on Population, 2019).

*** Sex ratio, Age structure, Rural-urban population ratio, Location of the source of drinking water, Education level by age and sex, Distribution of households by the condition of houses, Household size, Houseless households, Migrant population with work/employment as a reason for migration.

up of the data in standalone storage devices will be carried out. The findings of the study will be disseminated in the form of a Ph.D. thesis and publications. All the stakeholders will be duly acknowledged and informed about the findings of the study.

2.7. Statistical and machine learning software

The proposed study will be conducted using Free and open-source software (FOSS) and reproducible algorithms. The algorithms will be prepared and executed in the R environment using R version 4.0.3 (R Core Team, 2020) and above.

2.8. Project management

The coursework to undertake the present study by the principal investigator has been completed as part of the Ph.D. program under the guidance of an interdisciplinary Doctoral Advisory Committee in the last two years. Necessary permissions and clearances have also been obtained for the conduct of the study. Data collection and extraction is currently under process. The expected timeline of events for the present study is represented schematically in Fig. 2.

3. Expected outcomes

Value addition to the existing surveillance for dengue, by incorporating data sources from outside the health sector. The project will provide insights as to how we can strengthen dengue surveillance by using routine data as well as integrating multiple data sources from non-health sectors available in the country. The present study is expected to provide insights into the spatiotemporal patterns of dengue in the state of Punjab, India. The framework for routine data-based vector-borne disease forecasting models shall provide reproducible processes and algorithms in open-source platforms to be followed for the development of disease forecasting models using the data science approach.

4. Discussion

Health care in India is among the most dynamic and challenging sectors due to the constraints for human resources, inequitable access to health care services, and a reactive approach for essential health care among others (NITI Aayog, 2018). Government of India initiatives to achieve Universal Health Coverage in the country urges for the development of reproducible machine learning algorithms for fulfilling unmet health needs in the country. Though, enormous datasets are generated, reported, stored, and disseminated, initiatives and research projects for knowledge generation from the same are limited in the Indian subcontinent (Hung et al., 2020). There is an unmet need for analyzing the potential of routine data sources from multiple sectors for knowledge generation and development of evidence-informed decision-making process for disease prevention and control (NITI Aayog, 2018).

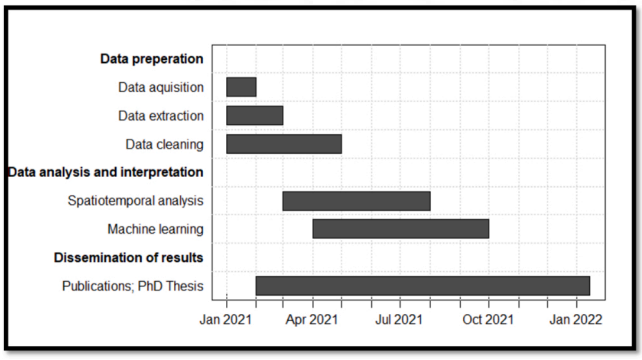


Fig. 2. Data collection and analysis plan.

Data availability from multisectoral initiatives accompanied by advances in data science is enabling the development of public health decision support tools and early warning systems across many regions of the globe (NASA, 2021). However. interlinking of datasets from multiple routine data sources provides both opportunities and challenges (Maastricht University, Netherlands and Künn, 2015). The datasets are available in varied formats such as HDF, NetCDF, excel, spatial multi polygons, and spatial rasters, among others. These raw datasets require pre-processing pipelines to integrate them at a defined spatial and temporal resolution for analysis. Further, the challenge of addressing changing administrative boundaries since Census 2011, missing value management, misplaced/ erroneous data entry issues, geocoding, and development of time- and space-wide datasets from satellite imagery datasets require the use of data extraction and imputation pipelines as an iterative process.

The present study includes NVBDCP line list data for dengue cases. Though the routine program datasets have received constant criticism for underreporting and data quality issues, it has been used for understanding disease patterns and to develop early warning systems in multiple regions across the globe (Louis et al., 2014). However, based on the data availability, varied timestamps have been used in disease modeling (annual (Stolerman et al., 2019), monthly (Zhang et al., 2016), weekly (Phanitchat et al., 2019), and daily (Titus Muurlink et al., 2018)). Also, a varied spectrum of case definitions have been used to explore associations of disease occurrence (suspected (Wangdi et al., 2018) / probable (Verma et al., 2018) / confirmed (Swain et al., 2019)).

The study includes the use of research-level satellite imagery datasets. The use of satellite imagery data for climatic and environmental variables have been recommended in low- and middle-income countries wherein the data availability from ground meteorological stations is sparse (Albarakat and Lakshmi, 2019; Sun et al., 2018). Satellite imagery datasets provide information on a finer spatial resolution as well as on regular time stamps. However, the satellite imagery includes noise in the data because of the aberrations from environmental conditions such as cloud covers. Inter-disciplinary research initiatives at global and national levels facilitate interdisciplinary research in public health epidemiology through the provision of pre-processed research-level datasets (NASA, 2021; National Remote Sensing Centre, 2020). These datasets are based on well underlying models that combine observations consistently to produce pre-processed datasets by domain expert teams. Further, the ground-truthing and validation studies have proved the high correlation of the satellite imagery data with ground stations (Albarakat and Lakshmi, 2019; Sun et al., 2018).

The present study envisages the exploration and development of reproducible algorithms and pipelines for the conduct of spatial, time series, and spatiotemporal analytics. Though knitted closely in understanding disease epidemiology, these approaches provide independent interpretations for the development of forecasting models. A “one size fits all” approach cannot fulfill requirements at multiple levels of health care. For example, on the one hand, a health manager may be required to undertake spatial analysis for identification of dengue clusters to design interventions for a large area with multiple administrative divisions, whereas on the other, a time series analysis for a limited geographic region. Task shifting, development of accountability and responsibility at multiple levels in health care, and other reforms for strengthening disease surveillance, prevention, and control can be brought by developing customizable algorithms in decision support tools to fulfill varied requirements.

Though the current study explores climatic, environmental, socio-demographic and spatiotemporal aspects of dengue occurrence, the current study has certain limitations. The limitations brought by lack of spatiotemporal routine data availability include lack of additional determinants for dengue occurrence and reporting such as various entomological, serological, behavioral, health system and transportation-related parameters. Future work in the development of such datasets and studies incorporating these parameters in the modeling process are

required to further enhance the robustness of scientific evidence. Lack of adequate spatial and temporal resolution in routine vector surveillance data hinders its application in the development of dengue models from low- and middle-income countries (Azil et al., 2011). The entomological vector surveillance data are often collected during outbreaks and by select research institutions resulting in patchy datasets. However, the association between dengue transmission by the poikilothermic *Aedes* mosquito and environmental variables has been established in the literature (Fan et al., 2014).

To conclude, the spatial and temporal patterns of a disease are inherent characteristics of routine health data. Also, data indicators related to disease risk profiling are generated at enormous volumes from non-health sectors. Understanding of Spatio-temporal patterns of dengue epidemiology and its association with predictor variables from routine data using a reproducible data science approach is an essential step towards creating an enabling environment for data-driven public health decision making.

Credit author statement

Gurpreet Singh: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing, Visualization, Project administration.

Biju Soman: Conceptualization, Methodology, Software, Supervision, Writing, Visualization, Project administration.

Availability of data and materials

All the datasets except dengue line listing data and spatial datasets are available in the public domain. The dengue line list data that support the findings of this study are available from the Directorate of Health Services, Government of Punjab, India and the spatial datasets are available from Punjab Remote Sensing center, Punjab, India but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of the Directorate of Health Services, Government of Punjab, India, or Punjab Remote Sensing center, Punjab, India as applicable.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

None.

Acknowledgments

We would like to acknowledge Dr. Manojkumar TK, Dr. Jeemon P, Dr. Shijulal Nelson Sathi, Dr. Srikant A, Dr. Gagandeep Singh Grover, and Dr. Arun Mitra P for their valuable inputs in the preparation of the study protocol.

References

van der Aalst, W.M.P., 2016. Process Mining: Data Science in Action, 2nd ed. Springer Berlin Heidelberg : Imprint: Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-49851-4. 2016.

Albarakat, R., Lakshmi, V., 2019. Comparison of normalized difference vegetation index derived from Landsat, MODIS, and AVHRR for the Mesopotamian marshes between 2002 and 2018. Remote Sens (Basel) 11, 1245. https://doi.org/10.3390/rs11101245.

Azil, A.H., Li, M., Williams, C.R., 2011. Dengue vector surveillance programs: a review of methodological diversity in some endemic and epidemic countries. Asia Pac. J. Public Health 23, 827–842. https://doi.org/10.1177/1010539511426595.

Banu, S., Hu, W., Hurst, C., Tong, S., 2011. Dengue transmission in the Asia-Pacific region: impact of climate change and socio-environmental factors. Trop. Med. Int. Health 16, 598–607. https://doi.org/10.1111/j.1365-3156.2011.02734.x.

Bouzillé, G., Poirier, C., Campillo-Gimenez, B., Aubert, M.-.L., Chabot, M., Chazard, E., et al., 2018. Leveraging hospital big data to monitor flu epidemics. Comput. Methods Programs Biomed. 154, 153–160. https://doi.org/10.1016/j.cmpb.2017.11.012.

Chae, S., Kwon, S., Lee, D., 2018. Predicting infectious disease using deep learning and big data. IJERPH 15, 1596. https://doi.org/10.3390/ijerph15081596.

Directorate General of Health Services, 2015. India. Joint Monitoring Mission Report.

Directorate General of Health Services, 2021a. Ministry of Health and Family Welfare, Government of India. DENGUE/DHF SITUATION IN INDIA. National Vector Borne Disease Control Programme https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=431&lid=3715 (accessed January 30, 2021).

Directorate General of Health Services, 2021b. Ministry of Health and Family Welfare, Government of India. National Vector Borne Disease Control Programme (NVBDCP) https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=405&lid=3681 (accessed January 30, 2021).

Fan, J., Wei, W., Bai, Z., Fan, C., Li, S., Liu, Q., et al., 2014. A Systematic review and meta-analysis of dengue risk with temperature change. IJERPH 12, 1–15. https://doi.org/10.3390/ijerph120100001.

Farrar, J., Manson, P., 2014. Manson's Tropical Diseases. 23. Elsevier Saunders, Edinburgh.

Government of Punjab, India, 2021. Know Punjab https://punjab.gov.in/know-punjab/ (accessed January 19, 2021).

Hung, Y.W., Hoxha, K., Irwin, B.R., Law, M.R., Grépin, K.A., 2020. Using routine health information data for research in low- and middle-income countries: a systematic review. BMC Health Serv. Res. 20, 790. https://doi.org/10.1186/s12913-020-05660-1.

Last, J.M., International Epidemiological Association, editors, 2001. A Dictionary of Epidemiology, 4th ed. Oxford University Press, New York.

Louis, V.R., Phalkey, R., Horstick, O., Ratanawong, P., Wilder-Smith, A., Tozan, Y., et al., 2014. Modeling tools for dengue risk mapping - a systematic review. Int. J. Health Geogr. 13, 50. https://doi.org/10.1186/1476-072X-13-50.

Maastricht University, Netherlands, Künn, S., 2015. The challenges of linking survey and administrative data. Izawol. https://doi.org/10.15185/izawol.214.

Minale, A.S., Alemu, K., 2018. Mapping malaria risk using geographic information systems and remote sensing: the case of Bahir Dar City. Ethiopia. Geospat Health 13. https://doi.org/10.4081/gh.2018.660.

Morin, C.W., Comrie, A.C., Ernst, K., 2013. Climate and Dengue Transmission: evidence and Implications. Environ. Health Perspect. 121, 1264–1272. https://doi.org/10.1289/ehp.1306556.

Murhekar, M.V., Kamaraj, P., Kumar, M.S., Khan, S.A., Allam, R.R., Barde, P., et al., 2019. Burden of dengue infection in India, 2017: a cross-sectional population based serosurvey. Lancet Glob. Health 7, e1065–e1073. https://doi.org/10.1016/S2214-109X(19)30250-5.

National Commission on Population, 2019. Population Projections for India and States 2011-2036. Ministry of Health and Family Welfare.

NASA, 2021. GPM Disease Inititative. Global Precipitation Measurement https://gpm.nasa.gov/applications/disease-initiative (accessed February 1, 2021).

National Centre for Disease Control, Directorate General of Health Services, 2021. Integrated Disease Surveillance Programme(IDSP) https://idsp.nic.in/index4.php?lang=1&level=0&linkid=313&lid=1592 (accessed January 30, 2021).

National Remote Sensing Centre, 2020. Bhuvan. Indian Geo-Platform of ISRO https://bhuvan-app3.nrsc.gov.in/data/download/# (accessed December 27, 2020).

NCBI, 2021. Data Science. MeSH https://www.ncbi.nlm.nih.gov/mesh/2028050 (accessed January 30, 2021).

NITI Aayog, 2018. National Strategy for Artificial Intelligence.

Ong, J., Liu, X., Rajarethinam, J., Kok, S.Y., Liang, S., Tang, C.S., et al., 2018. Mapping dengue risk in Singapore using Random Forest. PLoS Negl. Trop Dis. 12, e0006587 https://doi.org/10.1371/journal.pntd.0006587.

Ooi, E.-.E., Gubler, D.J., 2009. Global spread of epidemic dengue: the influence of environmental change. Fut. Virol. 4, 571–580. https://doi.org/10.2217/fvl.09.55.

Pei, S., Kandula, S., Yang, W., Shaman, J., 2018. Forecasting the spatial transmission of influenza in the United States. Proc. Natl Acad. Sci. USA 115, 2752–2757. https://doi.org/10.1073/pnas.1708856115.

Phanitchat, T., Zhao, B., Haque, U., Pientong, C., Ekalaksananan, T., Aromseree, S., et al., 2019. Spatial and temporal patterns of dengue incidence in northeastern Thailand 2006–2016. BMC Infect. Dis. 19, 743. https://doi.org/10.1186/s12879-019-4379-3.

R Core Team, 2020. R: A language and Environment for Statistical Computing. Foundation for Statistical Computing, Vienna, Austria.

Sánchez-González, G., Condé, R., Noguez Moreno, R., López Vázquez, P.C., 2018. Prediction of dengue outbreaks in Mexico based on entomological, meteorological and demographic data. PLoS ONE 13, e0196047. https://doi.org/10.1371/journal.pone.0196047.

Singh, G., Tilak, R., Kaushik, S.K., 2019. Bio-eco-social determinants of Aedes breeding in field practice area of a medical college in Pune, Maharashtra. Indian J. Public Health 63, 324. https://doi.org/10.4103/ijph.IJPH_296_18.

Stolerman, L.M., Maia, P.D., Kutz, J.N., 2019. Forecasting dengue fever in Brazil: an assessment of climate conditions. PLoS ONE 14, e0220106. https://doi.org/10.1371/journal.pone.0220106.

Sun, W., Sun, Y., Li, X., Wang, T., Wang, Y., Qiu, Q., et al., 2018. Evaluation and correction of GPM IMERG precipitation products over the capital circle in northeast China at multiple spatiotemporal scales. Adv. Meteorol. https://doi.org/10.1155/2018/4714173.

Swain, S., Bhatt, M., Pati, S., Soares Magalhaes, R.J., 2019. Distribution of and associated factors for dengue burden in the state of Odisha, India during 2010–2016. Infect. Dis. Poverty 8, 31. https://doi.org/10.1186/s40249-019-0541-9.

Titus Muurlink, O., Stephenson, P., Islam, M.Z., Taylor-Robinson, A.W, 2018. Long-term predictors of dengue outbreaks in Bangladesh: a data mining approach. Infect. Dis. Modell. 3, 322–330. https://doi.org/10.1016/j.idm.2018.11.004.

Vector-borne diseases. 2020. https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases (accessed January 19, 2021).

Verma, M., Kishore, K., Kumar, M., Sondh, A.R., Aggarwal, G., Kathirvel, S., 2018. Google search trends predicting disease outbreaks: an analysis from India. Healthc Inform Res 24, 300. https://doi.org/10.4258/hir.2018.24.4.300.

Volkova, S., Ayton, E., Porterfield, K., Corley, C.D., 2017. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. PLoS ONE 12, e0188941. https://doi.org/10.1371/journal.pone.0188941.

Wang, X., Tang, S., Wu, J., Xiao, Y., Cheke, R.A., 2019. A combination of climatic conditions determines major within-season dengue outbreaks in Guangdong Province. China. Parasites Vect. 12, 45. https://doi.org/10.1186/s13071-019-3295-0.

Wangdi, K., Clements, A.C.A., Du, T., Nery, S.V., 2018. Spatial and temporal patterns of dengue infections in Timor-Leste, 2005–2013. Parasites Vect. 11 (9) https://doi.org/10.1186/s13071-017-2588-4.

Withanage, G.P., Viswakula, S.D., Nilmini Silva Gunawardena, Y.I., Hapugoda, M.D, 2018. A forecasting model for dengue incidence in the District of Gampaha. Sri Lanka. Parasites Vectors 11, 262. https://doi.org/10.1186/s13071-018-2828-2.

World Health Organization, 2020. Dengue and Severe Dengue https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue (accessed January 29, 2021).

Xu, J., Xu, K., Li, Z., Meng, F., Tu, T., Xu, L., et al., 2020a. Forecast of dengue cases in 20 Chinese cities based on the deep learning method. Int. J. Environ. Res. Public Health 17, E453. https://doi.org/10.3390/ijerph17020453.

Xu, Z., Bambrick, H., Yakob, L., Devine, G., Frentiu, F.D., Villanueva Salazar, F., et al., 2020b. High relative humidity might trigger the occurrence of the second seasonal peak of dengue in the Philippines. Sci. Total Environ. 708, 134849 https://doi.org/10.1016/j.scitotenv.2019.134849.

Yuan, M., Boston-Fisher, N., Luo, Y., Verma, A., Buckeridge, D.L., 2019. A systematic review of aberration detection algorithms used in public health surveillance. J. Biomed. Inform. 94, 103181 https://doi.org/10.1016/j.jbi.2019.103181.

Zambrana, J.V., F, Bustos Carrillo, Burger-Calderon, R., Collado, D., Sanchez, N., Ojeda, S., et al., 2018. Seroprevalence, risk factor, and spatial analyses of Zika virus infection after the 2016 epidemic in Managua. Nicaragua. Proc Natl Acad Sci USA 115, 9294–9299. https://doi.org/10.1073/pnas.1804672115.

Zhang, J., Nawata, K., 2018. Multi-step prediction for influenza outbreak by an adjusted long short-term memory. Epidemiol. Infect. 146, 809–816. https://doi.org/10.1017/S0950268818000705.

Zhang, Y., Wang, T., Liu, K., Xia, Y., Lu, Y., Jing, Q., et al., 2016. Developing a time series predictive model for dengue in Zhongshan, China based on weather and guangzhou dengue surveillance data. PLoS Negl Trop Dis 10, e0004473. https://doi.org/10.1371/journal.pntd.0004473.

Zheng, L., Ren, H.-.Y., Shi, R.-.H., Lu, L., 2019. Spatiotemporal characteristics and primary influencing factors of typical dengue fever epidemics in China. Infect. Dis. Poverty 8, 24. https://doi.org/10.1186/s40249-019-0533-9.

Original Article

Development and Use of a Reproducible Framework for Spatiotemporal Climatic Risk Assessment and its Association with Decadal Trend of Dengue in India

Gurpreet Singh, Arun Mitra, Biju Soman

Achutha Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Thiruvananthapuram, Kerala, India

Abstract

Introduction: The study aimed to develop a reproducible, open-source, and scalable framework for extracting climate data from satellite imagery, understanding dengue's decadal trend in India, and estimating the relationship between dengue occurrence and climatic factors. **Materials and Methods:** A framework was developed in the Open Source Software, and it was empirically tested using reported annual dengue occurrence data in India during 2010–2019. Census 2011 and population projections were used to calculate incidence rates. Zonal statistics were performed to extract climate parameters. Correlation coefficients were calculated to estimate the relationship of dengue with the annual average of daily mean and minimum temperature and rainy days. **Results:** Total 818,973 dengue cases were reported from India, with median annual incidence of 6.57 per lakh population; it was high in 2019 and 2017 (11.80 and 11.55 per lakh) and the Southern region (8.18 per lakh). The highest median annual dengue incidence was observed in Punjab (24.49 per lakh). Daily climatic data were extracted from 1164 coordinate locations across the country for the decadal period (4,249,734 observations). The annual average of daily temperature and rainy days positively correlated with dengue in India ($r = 0.31$ and 0.06, at $P < 0.01$ and 0.30, respectively). **Conclusion:** The study provides a reproducible algorithm for bulk climatic data extraction from research-level satellite imagery. Infectious disease models can be used to understand disease epidemiology and strengthen disease surveillance in the country.

Keywords: Climate risk, dengue, public health, remote sensing, reproducible approach, satellite imagery, spatiotemporal

Introduction

Spatiotemporal and machine learning approaches are increasingly used to understand the epidemiology of infectious diseases.[1] The epidemiological understanding gained using these approaches has been instrumental in developing decision support tools, early warning systems, aberration detection algorithms, disease forecasting models, and evidence-informed public health decision-making.[2-4] Implementation of Integrated Health Information Portal, deregulation of geospatial data by Department of Science and Technology, National Digital Health Mission, and other digital health initiatives will generate high-resolution geocoded big data on health-related events in India in the coming years.[5-7] Existing routine datasets have also been used to understand micro-climatic determinants using algorithms that can extract spatiotemporal parameters associated with disease occurrence.[2]

The development of infectious disease models in low-and middle-income countries is faced with challenges of obtaining high-resolution data on climatic risk variation from on-ground meteorological stations. Global and National intersectoral initiatives provide satellite imagery-based Analysis Ready Datasets (ARDs) and global climatic models through multiple sources.[8-11] The use of these ARDs will enable public health managers and epidemiologists to obtain high-resolution climatic data,

Address of the correspondence: Dr. Biju Soman, Achutha Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Thiruvananthapuram, Kerala, India.
E-mail: bijusoman@sctimst.ac.in

Access this article online

Quick Response Code:

Website: www.ijcm.org.in

DOI: 10.4103/ijcm.ijcm_862_21

How to cite this article: Singh G, Mitra A, Soman B. Development and use of a reproducible framework for spatiotemporal climatic risk assessment and its association with decadal trend of dengue in India. Indian J Community Med 2022;47:50-4.

Received: 02-06-21, **Accepted:** 30-11-21, **Published:** 16-03-22

providing a future opportunity to strengthen existing disease surveillance.

Dengue is hyperendemic in India, and resultant economic losses have surpassed other vector-borne diseases.[12] The occurrence of dengue is critically determined by the microclimatic conditions.[13,14] Satellite imagery ARDs and preprocessed climatic models are routine data sources on microclimatic conditions which can be modeled for dengue analytics.[15] The incorporation of lagged climatic variables and spatial characteristics in such models establishes temporality as defined in Hill's criteria and adheres to Tobler's law in geography.

Satellite imagery ARDs are large datasets commonly available in Hierarchical Data Formats (HDF), network Common Data Form (NetCDF), and other data formats (Application Programming Interface [API] based).[9-11,16] Moderate Resolution Imaging Spectroradiometer provides ARDs in HDF; Integrated Multi-satellitE Retrievals for Global Precipitation Mission (IMERG) datasets, and Indian Meteorological Department (IMD) in NetCDF format; and Modern-Era Retrospective analysis for Research and Applications, Version 2, Meteorological and Oceanographic Satellite Data Archival Centre, Bhuvan web portal, and Open Government Data Platform India are API-based routine geospatial data sources. Handling large datasets in a reproducible environment increases the grade of evidence, reduces manual errors, and is computationally efficient.[17] Thus, the present study was conducted to explore and develop a reproducible framework for extracting spatiotemporal climatic risk parameters from satellite imagery ARDs, understand the decadal trend of dengue in India, and estimate the relationship between dengue occurrence and climatic factors in India.

Materials and Methods

Study design

The study was carried out in two phases. The first phase included exploring and developing a reproducible framework for research-level satellite imagery bulk preprocessing. The second phase included ecological analysis of publicly available dengue occurrence data and climatic variables obtained using the developed framework.

Exploration and reproducible framework development

Algorithms provided by Level-1 and Atmosphere Archive and Distribution System Distributed Active Archive Center, IMD Gridded datasets archive, Global Precipitation Mission, R package archives, GitHub, and other code repositories were explored. Proprietary software-based algorithms and algorithms for platforms other than the R environment were excluded. Framework for HDF, NetCDF, and API-based satellite imagery ARDs extraction into analyzable tidy data formats was developed.

Secondary data sources

Annual state-wise dengue occurrence data for the decadal period from January 01, 2010 to December 31, 2019, was extracted from the National Health Profile reports and National Vector Borne Disease Control Programme, India website.[18,19] Population estimates from Census 2011 and population projections for the year 2012–2019 provided population denominators for calculating dengue incidence per lakh population.[20] Climatic variables (temperature (mean and minimum) and cumulative precipitation) for daily timestamps were extracted using the "*nasapower*" package.[21]

Data analysis and interpretation

The National-, regional-, and state-level decadal trend of dengue was calculated. For regional level analysis, the zonal councils as defined by the Ministry of Home Affairs were adopted.[22] Zonal statistics were performed to calculate climate parameters. Descriptive measures were calculated for climatic variables. Data visualization using the GIS environment in an open-source platform was carried out. Correlation coefficients were calculated to estimate the relationship of dengue with mean annual temperature and rainy days. A $P < 0.05$ was considered statistically significant. The framework development and statistical analysis were carried out using R version 4.0.3 (R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria).[23]

Ethics statement

The present study is part of a larger research project culminating in the Ph.D. program of the first author. Institutional Ethics Committee (IEC/IEC-1653; IEC Reg. No. ECR/189/Inst/KL/2013/RR-16) clearance obtained vide letter SCT/IEC/IEC-1653/DECEMBER-2020 dated 19/12/2020.

Results

Framework

The algorithm undertakes bulk data extraction of climate parameters for a multi-polygon from stored NetCDF/HDF files. The files should be downloaded according to the instructions given on respective websites and stored in a file directory. All the steps for data extraction are automated in the algorithm based on user inputs on the path of the directory where NetCDF/HDF files are stored. The researcher/user should provide identification of desired sub dataset, scale factor, and offsets if any. For API-based extraction, a local grid with spatial resolution as of the data source is constructed for data extraction. The framework for data extraction from satellite imagery ARDs (NetCDF, HDF, and API-based) can be accessed on the GitHub repository.

Epidemiological trend of dengue in India

During the decadal period, 8,18,973 dengue cases were reported with a mean (standard deviation) annual incidence of 6.36 (3.60) per lakh population. The median annual dengue incidence for India was 6.57 per lakh population. Nationally, dengue incidence was maximum in 2019 followed by 2017 (11.80 and 11.55 per lakh, respectively), and minimum dengue incidence was in 2011 (1.56 per lakh).

Regionally, the highest median annual dengue incidence was observed in the South, followed by the West, North, North East, Central, and East region (8.18, 8.05, 4.5, 1.89, 1.62, and 1.6 per lakh, respectively). Among the states, the highest median annual dengue incidence was observed in Punjab, Goa, Kerala, and Odisha (24.49, 14.41, 12.13, and 9.1 per lakh, respectively). The Union Territories showed higher dengue incidence rates with the highest median annual incidence reported from Dadar and Nagar Haveli (126.22 per lakh), followed by Puducherry (77.45 per lakh). The national capital, Delhi, reported a median annual incidence of 28.70 per lakh population. Lakshadweep was the only state/UT with zero reported cases during the decadal period. Further, among the states, the outbreak years, as indicated by unusually high (more than 50 per lakh) dengue incidence were reported from the states of Arunachal Pradesh (134 per lakh in 2015), Uttarakhand (95 per lakh in 2019), Sikkim (66 per lakh in 2019), Goa (64 per lakh in 2019), Himachal Pradesh (64 per lakh in 2018), Kerala (57 per lakh in 2017), and Punjab (52 per lakh in 2017). The highest dengue incidence among union territories was reported from Dadar and Nagar Haveli (921 per lakh in 2016 and 427 per lakh in 2017), followed by Puducherry (318 per lakh in 2017 and 274 per lakh in 2012).

Climatic trends in India

Daily climatic data were extracted from 1164 coordinate locations across the country for the decadal study period (4,249,734 observations). The regional summary of decadal temperature is represented in Figure 1. The West, South, Central, and East regions of the country were warmer (decadal mean temperature of 26.31, 26.22, 26.31, and 25.41°C respectively) compared to North and Northeast regions (decadal mean temperature of 18.71 and 19.47°C, respectively). The temperature variation was maximum in the North region (IQR 17.37) and minimum in the South (IQR 3.78). The highest decadal mean rainfall was present in the Northeast region, followed by the South and East regions (75.53, 67.04, and 62.66 mm, respectively).

Correlation between dengue occurrence and climatic variables

The correlation between climatic variables and dengue is represented in Table 1. The annual average daily mean temperature was positively correlated with dengue at the national level (r = 0.31, P < 0.01). At the regional level, the correlation between mean temperature and dengue was maximum in West, North, and Central regions (r = 0.43, 0.37, and 0.35, P = 0.02, < 0.01, and 0.13 respectively). The annual average of daily minimum temperature was significantly correlated with dengue in the East and Northeast regions (r = 0.33 and-0.32, P = 0.04 and < 0.01, respectively). The precipitation days were positively correlated with the dengue at the national level (r = 0.06, P = 0.30). At the regional level, the East and Northeast regions had a statistically significant relationship between precipitation days and dengue (r = 0.38 and 0.28, respectively, P = 0.02).

Discussion

The present study documents availability of high-resolution satellite imagery research-level datasets and provide a reproducible algorithm for bulk data extraction and preprocessing of these datasets. Availability of micro-climatic data enables the development of models for understanding knowledge gaps in infectious disease epidemiology.[1,13,14,24] Advances in technology and increasing geocoded health data generation provide a challenge and an opportunity for the growth of epidemiological theories. Digital healthcare epidemiology, as compared to conventional epidemiology, is based on routine unstructured big datasets and requires a data science approach.[25] Research with reproducible open-source algorithms facilitates understanding of the research pathways and enables future expansion of existing frameworks.[17,26]

Satellite remote sensing has increased manifold in the past few decades in technology and application potential. High-resolution and multi-frequency satellite sensors can capture data on multiple climatic and environmental parameters, among others.[27] A validation study of the IMERG rainfall dataset with IMD gridded data showed a correlation of + 0.88 in India.[28] It is also essential to understand that raw satellite imagery datasets have inherent data quality issues and require technical proficiency for preprocessing. Thus, the availability of research-level datasets from domain expert teams helps public health professionals and epidemiologists to estimate the spatiotemporal variation of risk factors in disease causation.

The decadal dengue trend in India showed an increase across the country. This may be attributed to an actual increase over the decadal period and enhanced diagnostics, surveillance, and reporting mechanisms in the country. The correlation of climatic factors was found to be varying across regions in the country. It may be attributed to the large geographical extent and presence of multiple climatic zones. Temperature between 16-30 degrees Celsius is optimal for dengue transmission.[29] Precipitation provides water habitat for immature stages in the mosquito life cycle; however, high precipitation leading to flushing of immature stages is likely to have a negative association with dengue occurrence. In a study carried out to assess climatic factors and dengue occurrence in Thailand, different climatic factors were found to be associated with dengue incidence in coastal areas and plains.[30] Further studies at a more granular level (district/sub-district) are required to understand micro-climatic risk variation and its association with dengue in India.

The limitations in the present study include the lack of availability of granular dengue occurrence data. Data with a higher spatial and temporal resolution of disease occurrence would have further enhanced the understanding of the spatiotemporal epidemiology of dengue and its microclimatic associations. Furthermore, higher resolution data is required to understand the variance in these associations as per topography. These were beyond the scope of the present study. The role

of bio-eco-social determinants on the association of climatic factors with dengue occurrence was not studied in the present study. Incorporation of the same will enable the development of forecasting models to strengthen disease surveillance. The strength of the present study was the novel approach of using satellite imagery data to estimate the association between climatic factors and decadal dengue trends at national, regional, and state levels in India and the ability of the reproducible algorithm to process 4.2 million observations capturing daily climatic variables over a decade in a reproducible manner. The algorithms developed can be utilized in understanding the epidemiology of diseases affected by climatic conditions. The algorithm, being open-source and scalable, can be expanded to include additional satellite datasets in the future.

Collaborative studies between health departments and academic institutions with granular dengue surveillance data need to be conducted for understanding micro-climatic associations of dengue. Further, additional covariates such as climatic, environmental, sociodemographic, behavioral, and health system characteristics should be incorporated to understand the complex interplay of factors associated with dengue transmission. This understanding will enable us to develop efficient disease prevention and control strategies in the country.

Conclusion

The present study documents and provides a reproducible, systematic algorithm for spatiotemporal climatic risk assessment using research-level satellite imagery datasets. Further, the study highlights heterogenous high dengue burden in the country associated with climatic factors. The data science approach for spatiotemporal modelling of dengue incorporating climatic variables has the potential to develop forecasting models for strengthening routine surveillance in the country.

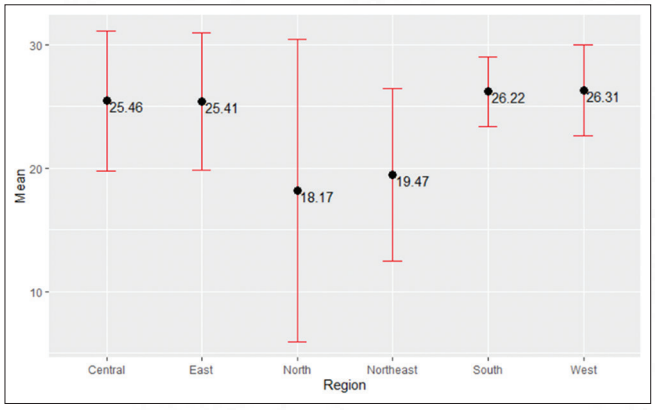


Figure 1: Region-wise decadal temperature distribution in India

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. An Overview of GeoAI Applications in Health and Healthcare | International Journal of Health Geographics | Full Text. Available from: https://ij-healthgeographics.biomedcentral.com/articles/10.1186/s12942-019-0171-2. [Last accessed on 2021 Feb 20].
2. Hung YW, Hoxha K, Irwin BR, Law MR, Grépin KA. Using routine health information data for research in low- and middle-income countries: A systematic review. BMC Health Serv Res 2020;20:790.
3. Carvajal TM, Viacrusis KM, Hernandez LF, Ho HT, Amalin DM, Watanabe K. Machine learning methods reveal the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila, Philippines. BMC Infect Dis 2018;18:183.
4. Shi Y, Liu X, Kok SY, Rajarethinam J, Liang S, Yap G, *et al.* Three-month real-time dengue forecast models: An early warning system for outbreak alerts and policy decision support in Singapore. Environ Health Perspect 2016;124:1369-75.
5. IHIP-Integrated Health Information Platform. Available from: https://idsp.nhp.gov.in/#!/. [Last accessed on 2021 Feb 20].
6. Department of Science and Technology. Guidelines for Acquiring and Producing Geospatial Data and Geospatial Data Services including Maps; DST F.No.SM/25/02/2020. Available from https://dst.gov.in/sites/default/files/Final%20Approved%20Guidelines%20on%20Geospatial%20Data.pdf [Last accessed on 2021 Feb 20].
7. NDHM. Available from: https://ndhm.gov.in/. [Last accessed on 2021 Feb 19].
8. Open Government Data (OGD) Platform India. Open Government Data (OGD) Platform India. Available from: https://data.gov.in/. [Last accessed on 2021 Feb 20].
9. IMERG: Integrated Multi-satellitE Retrievals for GPM | NASA Global Precipitation Measurement Mission. Available from: https://gpm.nasa.gov/data/imerg. [Last accessed on 2021 Feb 19].
10. MODIS Web. Available from: https://modis.gsfc.nasa.gov/data/dataprod/. [Last accessed on 2021 Feb 19].
11. National Remote Sensing Centre. Bhuvan. Indian Geo-Platform of

Table 1: Correlation between dengue occurrence and climatic factors

Region	Mean temperature (°C)		Minimum temperature (°C)		Precipitation (days)	
	Correlation coefficient (95% CI)	*P*	Correlation coefficient (95% CI)	*P*	Correlation coefficient (95% CI)	*P*
India	0.31 (0.20-0.41)	<0.01	−0.07 (−0.19-0.04)	0.22	0.06 (−0.05-0.18)	0.30
North	0.37 (0.17-0.55)	<0.01	0.14 (−0.08-0.35)	0.21	−0.04 (−0.26-0.18)	0.71
South	0.12 (−0.19-0.42)	0.44	0.06 (−0.25-0.36)	0.71	0.14 (−0.17-0.44)	0.37
East	0.19 (−0.14-0.49)	0.26	0.33 (0.0-0.6)	0.04	0.38 (0.05-0.64)	0.02
West	0.43 (0.09-0.69)	0.02	0.17 (−0.20-0.50)	0.37	−0.16 (−0.49-0.21)	0.39
Central	0.35 (−0.10-0.68)	0.13	0.03 (−0.42-0.46)	0.91	−0.38 (−0.7-0.07)	0.09
Northeast	0.13 (−0.12-0.36)	0.31	−0.32 (−0.53-−0.08)	<0.01	0.28 (0.03-0.49)	0.02

CI: Confidence interval

ISRO; 2020. Available from: https://bhuvan-app3.nrsc.gov.in/data/download/#. [Last accessed on 2020 Dec 27].
12. World Health Organization. Dengue and Severe Dengue; June 23, 2020. Available from: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue. [Last accessed on 2021 Jan 29].
13. Fan J, Wei W, Bai Z, Fan C, Li S, Liu Q, *et al.* A systematic review and meta-analysis of dengue risk with temperature change. Int J Environ Res Public Health 2014;12:1-15.
14. Morin CW, Comrie AC, Ernst K. Climate and dengue transmission: Evidence and implications. Environ Health Perspect 2013;121:1264-72.
15. Louis VR, Phalkey R, Horstick O, Ratanawong P, Wilder-Smith A, Tozan Y, *et al.* Modeling tools for dengue risk mapping – A systematic review. Int J Health Geogr 2014;13:50.
16. Rosenzweig C, Horton RM, Bader DA, Brown ME, DeYoung R, Dominguez O. *et al*. Enhancing climate resilience at NASA centers: A collaboration between science and stewardship. Bull. Amer. Meteorol. Soc., 2014;95:1351-63, doi:10.1175/BAMS-D-12-00169.1.
17. Peng RD. Reproducible research and Biostatistics. Biostatistics 2009;10:405-8.
18. National Health Profile: Central Bureau of Health Intelligence. Available from: https://www.cbhidghs.nic.in/index1.php?lang=1&level=1&sublinkid=75&lid=1135. [Last accessed on 2021 Feb 20].
19. Directorate General of Health Services, Ministry of Health and Family Welfare, Government of India. DENGUE/DHF SITUATION IN INDIA. National Vector Borne Disease Control Programme; 2021. Available from: https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=431&lid=3715. [Last accessed on 2021 Jan 30].
20. National Commission on Population. Population Projections for India and States 2011 – 2036. Ministry of Health and Family Welfare, Govt of India; 2019. Available from: https://nhm.gov.in/New_Updates_2018/Report_Population_Projection_2019.pdf. [Last accessed on 2021 Feb 20].
21. Sparks A. Nasapower: A NASA POWER global meteorology, surface solar energy and climatology data client for R. JOSS 2018;3:1035.
22. Ministry of Home Affairs, Government of India. Zonal Council. Ministry of Home Affairs; 2017. Available from: https://www.mha.gov.in/zonal-council. [Last accessed on 2021 Feb 18].
23. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/. [Last accessed on 2021 Feb 18].
24. Wang X, Tang S, Wu J, Xiao Y, Cheke RA. A combination of climatic conditions determines major within-season dengue outbreaks in Guangdong Province, China. Parasit Vectors 2019;12:45.
25. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol 2019;188:2222-39.
26. Hemant P. Reproducible Machine Learning. Medium, 2020. Available from: https://towardsdatascience.com/reproducible-machine-learning-cf1841606805. [Last accessed on 2021 Feb 20].
27. Manikiam B. Satellite based climate change study. Vayu Mandal 2015;41:9.
28. Kumar TV, Barbosa HA, Thakur MK, Paredes-Trejo F. Validation of Satellite (TMPA and IMERG) Rainfall Products with the IMD Gridded Data Sets over Monsoon Core Region of India. In: B. Rustamov R, ed. Satellite Information Classification and Interpretation. IntechOpen, 2019. DOI:10.5772/intechopen.84999.
29. Farrar J, Manson P, editors. Manson's Tropical Diseases. 23rd ed. Edinburgh: Elsevier Saunders; 2014.
30. Promprou, S, Jaroensutasinee, M & Jaroensutasinee, K. (2005). Climatic Factors Affecting Dengue Haemorrhagic Fever Incidence in Southern Thailand. WHO Regional Office for South-East Asia. Available from https://apps.who.int/iris/handle/10665/164135 [Last accessed on 2021 Feb 20].

A SYSTEMATIC APPROACH TO CLEANING ROUTINE HEALTH SURVEILLANCE DATASETS: AN ILLUSTRATION USING NATIONAL VECTOR-BORNE DISEASE CONTROL PROGRAMME DATA OF PUNJAB, INDIA

Gurpreet Singh, Sree Chitra Tirunal Institute for Medical Sciences and Technology, India, drgurpreet.md.afmc@gmail.com

Biju Soman, Sree Chitra Tirunal Institute for Medical Sciences and Technology, India, bijusoman@sctimst.ac.in

Arun Mitra, Sree Chitra Tirunal Institute for Medical Sciences and Technology, India, arunmitra2003@gmail.com

Abstract. Advances in ICT4D and data science facilitate systematic, reproducible, and scalable data cleaning for strengthening routine health information systems. A logic model for data cleaning was used and it included an algorithm for screening, diagnosis, and editing datasets in a rule-based, interactive, and semi-automated manner. Apriori computational workflows and operational definitions were prepared. Model performance was illustrated using the dengue line-list of the National Vector Borne Disease Control Programme, Punjab, India from 01 January 2015 to 31 December 2019. Cleaning and imputation for an estimated date were successful for 96.1% and 98.9% records for the year 2015 and 2016 respectively, and for all cases in the year 2017, 2018, and 2019. Information for age and sex was cleaned and extracted for more than 98.4% and 99.4% records. The logic model application resulted in the development of an analysis-ready dataset that can be used to understand spatiotemporal epidemiology and facilitate data-based public health decision making.

Keywords. Routine data, Data Science, Data cleaning, Reproducible algorithm, open-source software.

1. INTRODUCTION

Routine Health Information Systems (RHIS) includes data that is collected at regular intervals from multiple health facilities including community-level public health centers, public and private hospitals, and other healthcare institutions (MEASURE Evaluation, 2021). These datasets provide information on health status, health services, and resources available for improving the health of populations. The strengthening of RHIS has emerged as a global as well as national agenda in numerous countries for data-driven decision-making. The processes involved in RHIS strengthening are thus looked at from a broader perspective beyond the data collection and entry processes. Harrison et al. suggest five pillars that form the basis of the simplified theory of change in strengthening routine health surveillance data for decision making: governance, people, tools, processes, and evidence (Harrison et al., 2020). The framework provided by World Health Organization to strengthen health systems includes health information as one among the identified attributes of a health system (World Health Organization, 2007). Also, the current existing initiatives such as the "Performance of Routine Information System Management" (PRISM) framework suggested by the Measure evaluation study group addresses many of the current challenges for improving data quality through the data life cycle. Some of the aspects involved in the data life cycle as suggested in the PRISM framework for evaluation of routine health information systems include behavioral challenges, environmental challenges, organizational challenges, and technological challenges (MEASURE Evaluation, 2021).

Good quality data is paramount to the success of health information systems. Generally, data is considered high-quality if it is "fit for [its] intended uses in operations, decision making and planning while representing the real-world constructs it" (Fadahunsi et al., 2019). Data quality of routine health information systems has been a subject of extensive research over the years. Availability of good quality data at timely intervals is critical to data-based public health decision-making (AbouZahr & Boerma, 2005). In addition to social, economic, political, and local contextual factors, multiple factors have been identified at each stage of the data life cycle which affect data quality. The quality of data collected is largely influenced by the level of work engagement, training, and perceived-self efficacy of the individual collecting data. Health system-related factors such as multiple communication channels, increasing variables for data entry, limited health infrastructure, and frequent changes in reporting formats are known challenges to good quality data (Aiga et al., 2008; Glèlè Ahanhanzo et al., 2014).

Advances in Information and Digital Technologies and data science approaches have potential in cleaning and extraction of information from routine large datasets. Routine health information datasets are prepared primarily for administrative and programmatic use. As a result, the data quality standards laid for monitoring data elements are thus bound to be defined differently when compared to those required for research-level datasets. This brings forward the need for data cleaning measures on raw routine datasets before use for research purposes. Inability to identify data anomalies efficiently leads to loss of information, high missing values, and inaccurate outcomes (Maïga et al., 2019; Van den Broeck et al., 2005). A systematic approach to data cleaning is recommended along with transparent documentation, however, there is a dearth of studies that explicitly disclose the steps followed and anomalies detected and corrected during data cleaning (Maina et al., 2017; Wilhelm et al., 2019). Further, it is essential to understand data cleaning as a systematic process rather than a one-time activity. The importance of data cleaning in the data lifecycle is crucial as the resultant data's quality would not only determine the robustness and generalizability but also allow for data linkage and sensible extrapolation of the study findings (Gesicho et al., 2020; Phan et al., 2020; Randall et al., 2013; Van den Broeck et al., 2005). Adopting a systematic approach to data cleaning would enable the researcher to find anomalies more efficiently and allow for reproducibility and transparency of the data lifecycle (Huebner et al., 2016).

The implications of open-source algorithms using technological advances on the future public health landscape are enormous. The volume of the data that flows through a health system is enormous and ever-increasing. Studies have documented that the data volume in the digital universe is doubling every two years (Oracle India, 2021). Further, data integrity and data consistency have been raised by many in context to the routine health information system (Smeets et al., 2011). Though a lot of light has been shed on data quality assurance and data quality control, these principles are yet to be translated into practice, especially in low-and-middle-income country settings including India. Evidence-informed data-based real-time decision-making by health program managers and data users require efficient data cleaning processes to extract information and knowledge from data. Manually, this process is not standardized, time and resource-intensive, and is often faced with manual omissions and commissions. The development of reproducible algorithms will enable efficient data cleaning on one hand and will provide solutions to numerous challenges which country is facing in terms of estimating the real-time burden of diseases, capacity building, rapid public health decision making, and thus enhanced prevention and control of diseases.

The use of open-source and reproducible algorithms will enable the generation of semi-automated mechanisms for data cleaning and provide transparency to the cleaning process followed. As it is important to study attributes related to the decision-makers such as data use culture, personal believes, and power relations in the organization to strengthen information systems, at the same time, it is prudent to look at the technological challenges when dealing with routine health information systems in the 21st century. The data science approach is useful in achieving this humongous task efficiently and scientifically. This is especially relevant as much of the data collected through the routine health information system currently is in the digital format. The data science approach also incorporates accountability and transparency which are now being realized as

key issues when it comes to the use of health information. Reproducible algorithms can be used for revealing patterns of disease and transform health-related data for public health decision-making.

National Vector Borne Disease Control Programme (NVBDCP) is the nodal program for the prevention and control of vector-borne diseases in India (NVBDCP, 2021). Routine surveillance of vector-borne diseases is being carried out and data is generated from multiple health facilities. Dengue is a notifiable disease in the state of Punjab, India, and line listing of lab-confirmed cases are prepared for detailed epidemiological investigations, administrative requirements, and as decision support for the institution of preventive and control measures. Use of these routine health surveillance datasets to understand spatiotemporal patterns of dengue and linking with routine data from non-health sectors to understand determinants (climatic, environmental, socio-demography, health systems, etc.) will enable an in-depth understanding of dengue situation and development of disease forecasting models. This will strengthen existing surveillance mechanisms, and thus improve the health of the populations. However, for cross-linking of datasets, and the conduct of data analytics for knowledge generation, it is essential to clean the datasets in a manner that analysis-ready datasets are prepared from raw data without losing information. Thus, the present study was conducted to develop a rule-based reproducible and scalable logic model for cleaning routine health surveillance data in India using NVBDCP, Punjab program data as an illustrative example.

2. MATERIAL AND METHODS

2.1. **Data source.** Routine health care surveillance data provided by National Vector Borne Disease Control Programme, Directorate of Health Services, Government of Punjab, India. The datasets are composed of line listing data of lab-confirmed Dengue cases in the state from 01 January 2015 to 31 December 2019.

2.2. **Study variables.** The variables extracted from the routine data line list included information on the age of the patient, gender, place of occurrence, type of test performed, testing facility, and dates of testing, reporting, outpatient consultation, admission, and discharge.

2.3. **Framework.** The present study was conducted using the framework provided by Broeck et al. for data cleaning as a process (Van den Broeck et al., 2005). According to the framework, a data cleaning process is integral to all the components of a study process viz. study designing, data collection, data transformation, data extraction, data transfers, data exploration, and data analysis. The data cleaning process is a logical sequence of screening, diagnosing, and editing. The datasets are screened for anomalies and diagnosed to determine whether the anomaly is a true normal value, true extreme, an error, or undiagnosed with available data. This is followed by editing the data values by correction, deletion, or leaving as unchanged.

2.4. **Study design.** The data science approach was used for screening, diagnosis, and editing of raw datasets within the broad framework stated above. A reproducible algorithm was prepared which included a query code for screening, check code for diagnosis, and correction code for editing of data. Apriori definitions for valid data values expected range, data type, outliers, and data entry anomalies were created. The raw datasets were then systematically screened using the prepared algorithm for validity, presence of additional information, inconsistencies, strange patterns, and misplaced data in the line list. Once identified, the observed value, expected value, and neighborhood values were compared to confirm the presence of data issues that can be cleaned using the algorithm. All the data anomalies which were diagnosed to be due to apriori definitions were cleaned and data was extracted using an automated algorithm. The data anomalies wherein strange patterns were identified but had implicit valid values, a manual correction was carried out. In case of failure to obtain any valid value, the respective data cell was considered to be missing.

2.5. **Logic model**. A schematic representation of the logic model used for the data cleaning process is represented in Figure 1. The model imports the dataset and applies a rule-based, interactive algorithm to develop tidy data. The algorithm developed enquires about multiple possible anomalies in the dataset which can be cleaned in a semi-automated manner to extract information, and thus avoid loss of information for the respective variables. Each inquiry was based on the

detection of string patterns in each reported case followed by an automated correction code on confirmation. The algorithm had subsets of inquiries for each variable and was run on all the cells in a phased manner for the data cleaning process. Data values were standardized by importing as text variables and creating a tidy string variable as the first step of the data cleaning process.

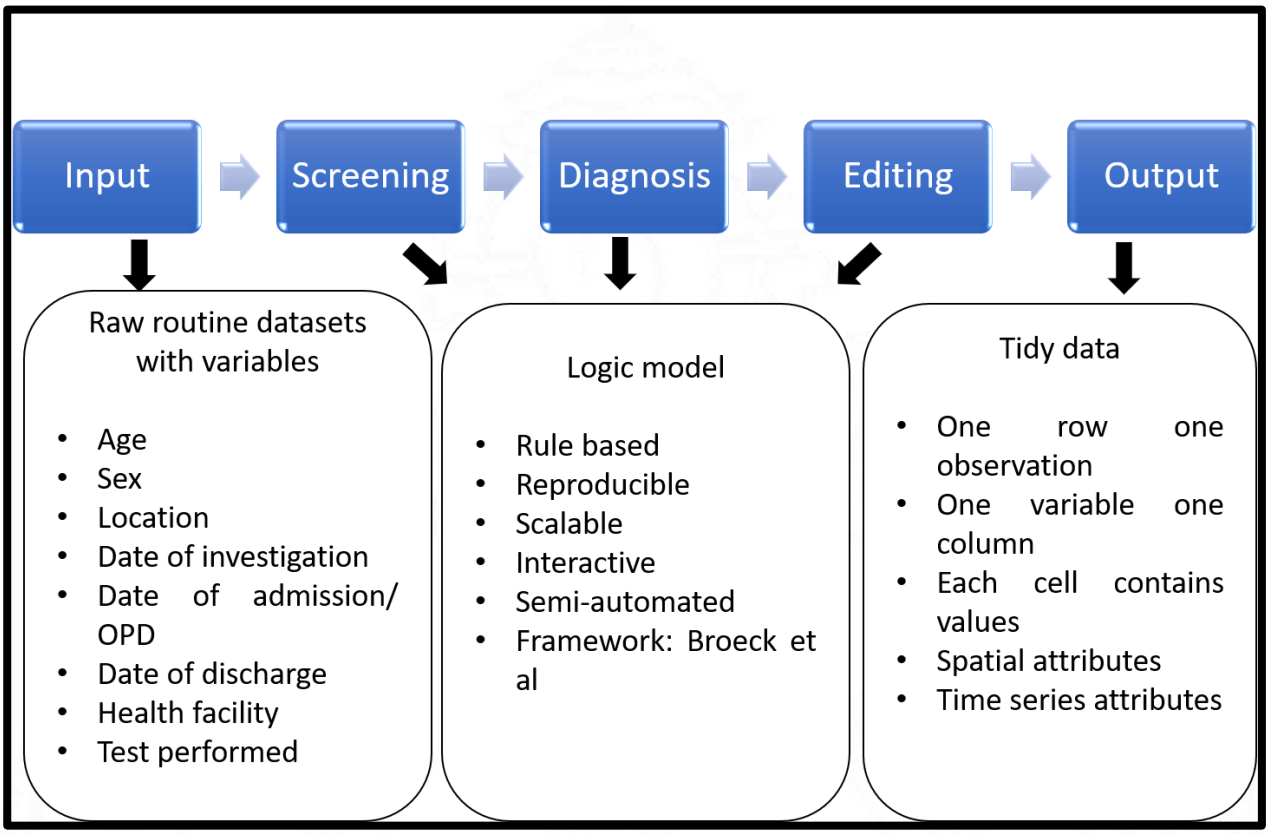


Figure 1 Schematic representation of the data cleaning process.

2.6. **Operational definitions and computational workflows.**

2.6.1. **Data cleaning process for date variables.** Date information can be analyzed when the date variable is present in a standard date format (e.g. ISO format). The date values were categorized into two types viz "excel-numeric" and "as-typed". Excel-numeric dates are values were defined as values with five characters, all digits, and no separator between digits. As-typed format values varied from a minimum length of 4 (e.g.: "2918" can represent for 02 September 2018) and had variations resulting from the field data entry personnel preferences (e.g.: "04 Jan 2020", "04-01-2020", "04/1/20", etc.). All the date values were read as string/text/character variables and standardized by removing all punctuations, separators, and whitespace. Each data value was then screened for the format of the date variable. In the case of the "excel-numeric" format, the date value was extracted by calculating the number of days since 01 January 1990. In the case of the "as typed" format, the value was screened for anomalies and data editing was carried out using the algorithm. All date values were transformed to the "ddmmyy" format for data extraction. Further, for missing data, data imputation by addition of mean days to testing from date of admission/OPD and subtracting mean days from discharge was carried out to estimate the date of testing.

2.6.2. **Data cleaning process to extract age-related information.** Age in analyzable format was defined as a numeric variable between 0 to 120 years. The expected column containing age details was imported as a text/ string/ character variable. All cells were screened for the presence of digit and character values. The values with the presence of non-digit characters were screened for the presence of valid digit values and data editing was carried out based on the findings. The alternate

columns (e.g. sex details) which are likely to contain misplaced values for missing data cells in the age column were screened in a phased manner for enhancing the data extraction process.

2.6.3. **Data cleaning process to extract sex details**. The sex variable was defined as a factor variable with three levels viz. Male, Female, and Transgenders. All the cells were screened for the presence of non-case-sensitive keywords including "Male", "Female", "M" without "F", "F", "Child", "Transgender", and "TG". Data values containing digit characters were cleaned by deletion of digits and standardizing character values to lower case, removal of punctuations, and whitespaces. To obtain data on missing values, other columns were screened in a phased manner using keywords.

2.6.4. **Data cleaning process to extract location details**. The address variable was defined as a string/ text variable containing information related to the district, sub-district/ block, city, village, and town details. To extract location details, the addresses were standardized before bulk geocoding. Bulk geocoding was done using google API client services. The location details were extracted from the raw address using a reference list of blocks, cities, towns, and villages adapted from Census 2011 datasets. Information related to the patient/ caretaker was found to be present in raw address datasets, especially for the children during exploratory data analysis. The same was removed from the raw addresses through regular expressions-based text mining approaches.

2.6.5. **Data anonymization**. All the identifiers such as name and contact details were removed from the dataset. R package. *Epitrix* was used to generate anonymized data using the "scrypt" algorithm.

2.7. **Software.** All the data cleaning algorithms were prepared and executed in R software (R Core Team, 2020) using *tidyverse, stringr, lubridate, ggmaps, and epitrix* packages.

2.8. **Ethics statement**. The present study is part of a larger research project culminating in the Ph.D. program of the first author. Institutional Ethics Committee (IEC/IEC-1653; IEC Reg. No. ECR/189/Inst/KL/2013/RR-16) clearance obtained vide letter SCT/IEC/IEC-1653/DECEMBER-2020 dated 19/12/2020. Permission for use of program data has been obtained from the Directorate of Health Services, Government of Punjab, India.

3. RESULTS

The algorithm was executed for line listing data of 64,688 lab-confirmed dengue cases reported from the state during the study period. Logic algorithm for date extraction with screening results, automated cleaning process using the algorithm, and manual corrections are represented in Table 1. A total of 2,04,985 cells expected to contain date values were screened. The excel-numeric format was found in 42,902 cell values. Among 1,31,931 values identified by the screening algorithm, 1,31,569 (99.72%) were cleaned using the apriori cleaning codes (automated) after confirmation of screening results, and 362 (0.27%) values required manual correction.

Logic algorithm	**Rationale**	**Automated data cleaning process**	**Screening results**	**Automated**	**Manual correction**
Total values screened	204985				
Date	To covert string variable to date, same pattern should be present across the cells for correct parsing. In this pipeline, all the different date formats are converted to ddmmyyyy format for standardization of pattern and thus correct parsing of dates.				
Presence of non-digit characters	Identify date specification using month values such as Jan, Feb, etc-	Remove all other non-digit characters.	41747	41747	0

Five-digit character values not starting with excel numeric date digit format and last two digits in yy format	All five-digit dmmyy/ddmyy format values will end in yy format.	Replace last two digits with yyyy year format.	5627	5627	0
Five-digit character values numerically outside the range of excel format dates and last two digits in yy format	Remaining five digit character values in as-typed format will be outside the range of number of days since 1899-12-30 to start and end days of the specified year.	Replace the last two digits with yyyy year format.	202	202	0
Five-digit character values numerically outside the range of excel format dates and not ending in yy format	Erroneous data values	Deletion.	257	228	29
Five-digit character values numerically within the range of excel format dates	Five-digit character values in excel format will be within the range of number of days since 1899-12-30 for the specified year	Excel format date extraction for five-digit values.	42902	42902	0
Any value with length more than eight characters.	Maximum length of dates in ddmmyyyy format is eight	Deletion.	565	512	53
Eight-digit character values					
Values not ending with yyyy format of the specified year	All eight-digit character dates should end with yyyy year format	Replace the last four digits with yyyy year format.	168	119	49
Month location holding a value greater than 12.	All eight-digit character dates should be ddmmyyyy format for parsing dates	Convert to ddmmyyyy format from mmddyyyy format.	81	74	7
Seven-digit character values					
Values ending digit as 1.	Mention of NS-1 positive along with dates in the raw data introduces error	Remove 1 from the last position.	443	438	5
Values not ending in yyyy format.	All seven-digit dates ending in yyyy format	Replace last four digits by 2019.	157	72	85
Values starting with 311, 211, or 111, and ending in the yyyy year format	Dates in seven-digit character have similar pattern on occurrence in month of January and November	If the date is in December or November, no changes are required. Else, replace the value with ddmmyyyy format.	931	931	0
Values with first two digits equal to	1st, 2nd, and 3rd of every month till September and 10th , 20th and 30th have	If the dates are 1st, 2nd or 3rd, (dmmyyyy) no changes required. Else	1181	1160	21

10 and ending in yyyy format.	same pattern in seven digit character format.	(ddmyyyy) insert a zero in the third location to create ddmmyyyy format.			
Values starting with zero.	Any seven-digit characters value starting cannot start with zero for parsing dates.	If the value is in ddmyyyy format, insert a zero at the third place.	38	36	2
Values ending in yyyy year format and numerically second and third location value is less than or equal to 12.	In seven-digit character values, from January to September, till 9th of every month (9122019), dates are written in dmmyyyy format. Then, from 11th to 31st of every month, dates are written in ddmyyyy format.	Insert a zero at first position to convert the value into ddmmyyyy format.	5447	5447	0
Remaining seven-digit character values	-	For values in ddmyyyy format, insert a zero at third location to convert it into ddmmyyyy format.	6900	6887	13
Six-digit character values					
Values starting with dmyyyy format.	Six-digit dates can be parsed and converted to date format when it is in ddmmyy format. In case of dmyyyy format or yyyydm format there will be error to parse.	Convert yyyydm into ddmmyyyy format.	395	389	6
Values which do not have last two digits as yy format.	All six-digit character values for a specified year should end in yy format	Replace last two digits with yy format.	617	529	88
Values which are ending in yy format, month location value is less than or equal to 12, and the date location value is less than or equal to 31	Six-digit character values in ddmmyy format should be converted into ddmmyyyy format for similar pattern across dates for easy parsing at a later stage	Replace last two digits with yyyy format.	23313	23310	3
Four-digit character values					
Values with yyyy format	Four-digit character values should be in dmyy format	Convert to ddmmyyyy format or manual correction	0	0	0
Values not ending with yy format	Values with ddmm format	Convert to ddmmyyyy format or manual correction	28	27	1

Remaining four-digit character values		All values in dmyy format to be converted to ddmmyyyy format by inserting additional zeros for day and month, and replacing yy with yyyy format, or manual correction	307	307	0
Values with three and less digit characters		Deletion/ Manual correction	625	625	0
Total			**131931**	**131569 (99.72%)**	**362 (0.27%)**

Table 1 Logic model characteristics and performance for date extraction.

Data extraction details for date, age, and sex variables from the dataset using the logic model are represented in **Figure 2**. The algorithm was able to clean and compute the estimated date of testing for 96.1% and 98.9% of observations for the year 2015 and 2016 respectively, and for all cases in the year 2017, 2018, and 2019. Age details were extracted maximum for the year 2017 and 2019 (99.4%) and minimum for the year 2015 (98.4%). Information on the sex of the patient was available for more than 99 percent across the study period, and location details were available for all the reported cases during the study period.

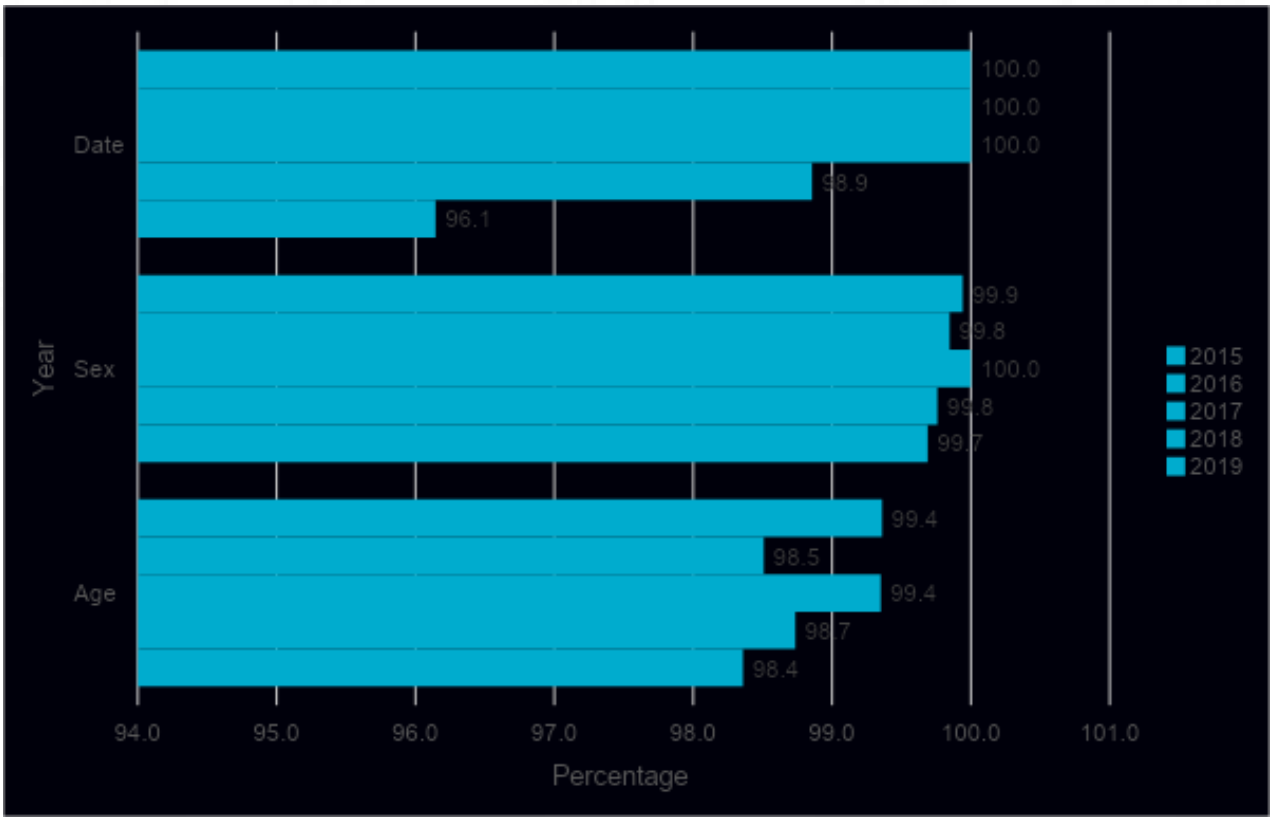


Figure 2 Data extraction summary for age, gender, and estimated date of testing.

4. DISCUSSION

The present study documents a systematic and reproducible logic model for data cleaning of routine health surveillance datasets in India. Though the systematic approach for data cleaning has been documented earlier on routine health information datasets (Gesicho et al., 2020; Maina et al., 2017; Phan et al., 2020), this study is novel in its application and illustration on routine program level dataset in India. Also, the present study was based on a data science approach that is increasingly being used for data analysis in epidemiology, but its utility in the development of reproducible and scalable data cleaning models has limited documentation. A recently published systematic review looking at the strategies applied in research articles to counter the issues of RHIS data quality across low- and middle-income countries suggest that majority of the studies that used RHIS data neither described the extent of the quality issues nor the steps they took to overcome them (Hung et al., 2020). The logic model developed in this study is expected to provide a practical strategy to clean routine health information program datasets in India resulting in strengthening of data quality for information and knowledge generation in the decision-making process as well as for research purposes.

The algorithm developed screened the data for date variables in a logical systematic approach. Among multiple variables present in the datasets, the timeline of disease occurrence is of utmost importance when analysis for disease patterns and model development is considered. Time series analysis is the most common analytical method followed by geostatistical analysis in routine data analytic studies (Hung et al., 2020). However, the dates are entered in varied formats in routine health information systems. This may be attributed to the use of basic data entry platforms such as Microsoft excel in the system. Though it suffices the "intended use" as defined for good quality data for day-to-day performance within the existing system, digital transformation of health care surveillance can be achieved by incorporating advancements in data handling and management technologies. Engagement of both data producers and users, identification of information needs, capacity building for data use at multiple levels, strengthening of data use and demand infrastructure are recommended measures for enhancement of data use context in health care systems (Nutley & Reynolds, 2013).

The present study used a rule-based semi-automated logic algorithm for data cleaning. Data cleaning approaches commonly used are broadly classified as logic-based and quantitative approaches. The use of Machine Learning and Artificial Intelligence based automated data cleaning workflows are largely based on the metadata of datasets. The semi-automated approach was chosen in the present context as it allows the user to understand the data along with the cleaning process in an iterative manner. The routine data currently in the country can be considered as digitalized as compared to the process of digital transformation wherein open data standards and metadata are inherent in database management systems. Initiatives for such database systems are required to enable the adoption of automated data cleaning workflows.

The presence of missing data values for the selected variables was found to be lower in the present study. This is in contrast to reported data missingness percentages in previous studies using routine datasets. This may be attributed to the type of variables selected and their perceived importance in the primary data use process. The line listing datasets prepared in the NVBDCP program include details on a limited number of variables that are considered essential for decision-making in the program. Further, the data values for a specified variable which seemed missing were found to be more commonly misplaced in the dataset. As a result, if alternate columns which are likely to hold information are not processed, the rate of missing data will be higher.

The research dissemination and uptake in health services require a collaborative approach between decision-makers and researchers to optimally utilize the advancements in information and digital technologies in health care. Data availability of program data for research purposes is required. Studies have proved that with increasing use of routine health data in decision making as well as for

research purposes creates a self-perpetuating milieu in the data environment leading to improved data quality and strengthening of health systems.

Study limitations. Understanding the reasons behind the data anomalies present in the routine datasets is a critical factor to guide interventions to improve data quality. However, its understanding was beyond the scope of the present study. The algorithm developed in the present study was based on a single disease dataset from the national vector-borne disease control program. Its application in other diseases and program datasets may require additional screening mechanisms on one hand and may not require some screening steps on the other. Future studies on the application of the algorithm for external generalizability will establish the robustness of the algorithm for larger use. Similarly, limited variables required for the present project were explored in the present study, however, being scalable, algorithms for additional variables as required by health program managers and for research purposes can be incorporated into the model.

The strength of the present study includes the use of a reproducible and scalable logic algorithm for data preprocessing of routine health surveillance data. This will enable the researchers to start looking at available routine datasets. The resulting dataset can be used to understand the Spatio-temporal epidemiology of diseases. Good quality data can be used to develop forecasting models which can complement existing surveillance mechanisms and reduce disease-related burden in the populations. The scalability of algorithms prepared in open-source software provides enormous potential for application to routine datasets for other diseases and for geographical regions with similar challenges globally. However, the conceptualization of development in ICT4D involves careful understanding of the research context within the broader goals for sustainable development. The institutionalization mechanisms for outcomes of ICT4D research will require ingrained perspectives related to the dimensions and theories of change for development (Zheng et al., 2018).

5. CONCLUSION

Data quality of routine health information systems can be strengthened using systematic, reproducible algorithms for data cleaning in open-source software. The algorithm in the present study was semi-automated and based on routine health surveillance data in India. It resulted in the development of a research-level dataset that can be analyzed and interlinked with data from non-health sectors, thus illuminating one of the key contributions of data science to public health systems. Being scalable, the implication of this information and digital technology in health systems and digital epidemiology is enormous. The logic model can be expanded for additional variables according to the health system and research needs in the future.

REFERENCES AND CITATIONS

AbouZahr, C., & Boerma, T. (2005). Health information systems: The foundations of public health. Bulletin of the World Health Organization, 83(8), 578–583. https://doi.org//S0042-96862005000800010

Aiga, H., Kuroiwa, C., Takizawa, I., & Yamagata, R. (2008). The reality of health information systems: Challenges for standardization. Bioscience Trends, 2(1), 5–9.

NVBDCP. (2021). DENGUE/DHF SITUATION IN INDIA. National Vector Borne Disease Control Programme. https://nvbdcp.gov.in/index4.php?lang=1&level=0&linkid=431&lid=3715

Fadahunsi, K. P., Akinlua, J. T., O'Connor, S., Wark, P. A., Gallagher, J., Carroll, C., Majeed, A., & O'Donoghue, J. (2019). Protocol for a systematic review and qualitative synthesis of information quality frameworks in eHealth. BMJ Open, 9(3), e024722. https://doi.org/10.1136/bmjopen-2018-024722

Glèlè Ahanhanzo, Y., Ouedraogo, L. T., Kpozèhouen, A., Coppieters, Y., Makoutodé, M., & Wilmet-Dramaix, M. (2014). Factors associated with data quality in the routine health information system of Benin. Archives of Public Health, 72(1), 25. https://doi.org/10.1186/2049-3258-72-25

Gesicho, M. B., Were, M. C., & Babic, A. (2020). Data cleaning process for HIV-indicator data extracted from DHIS2 national reporting system: A case study of Kenya. BMC Medical Informatics and Decision Making, 20(1), 293. https://doi.org/10.1186/s12911-020-01315-7

Harrison, K., Rahimi, N., & Carolina Danovaro-Holliday, M. (2020). Factors limiting data quality in the expanded programme on immunization in low and middle-income countries: A scoping review. Vaccine, 38(30), 4652–4663. https://doi.org/10.1016/j.vaccine.2020.02.091

Huebner, M., Vach, W., & le Cessie, S. (2016). A systematic approach to initial data analysis is good research practice. The Journal of Thoracic and Cardiovascular Surgery, 151(1), 25–27. https://doi.org/10.1016/j.jtcvs.2015.09.085

Hung, Y. W., Hoxha, K., Irwin, B. R., Law, M. R., & Grépin, K. A. (2020). Using routine health information data for research in low- and middle-income countries: A systematic review. BMC Health Services Research, 20(1), 790. https://doi.org/10.1186/s12913-020-05660-1

Maïga, A., Jiwani, S. S., Mutua, M. K., Porth, T. A., Taylor, C. M., Asiki, G., Melesse, D. Y., Day, C., Strong, K. L., Faye, C. M., Viswanathan, K., O'Neill, K. P., Amouzou, A., Pond, B. S., & Boerma, T. (2019). Generating statistics from health facility data: The state of routine health information systems in Eastern and Southern Africa. BMJ Global Health, 4(5), e001849. https://doi.org/10.1136/bmjgh-2019-001849

Maina, J. K., Macharia, P. M., Ouma, P. O., Snow, R. W., & Okiro, E. A. (2017). Coverage of routine reporting on malaria parasitological testing in Kenya, 2015–2016. Global Health Action, 10(1), 1413266. https://doi.org/10.1080/16549716.2017.1413266

MEASURE Evaluation. (2021). Routine Health Information Systems [Page]. https://www.measureevaluation.org/our-work/routine-health-information-systems

Nutley, T., & Reynolds, HeidiW. (2013). Improving the use of health data for health system strengthening. Global Health Action, 6(1), 20001. https://doi.org/10.3402/gha.v6i0.20001

Oracle India. (2021). What Is Big Data? https://www.oracle.com/in/big-data/what-is-big-data/

Phan, H. T. T., Borca, F., Cable, D., Batchelor, J., Davies, J. H., & Ennis, S. (2020). Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: Protocol and application to a large patient cohort. Scientific Reports, 10(1), 10164. https://doi.org/10.1038/s41598-020-66925-7

R Core Team. (2020). R: A language and environment for statistical computing (4.0.3) [Computer software]. Foundation for Statistical Computing. https://www.R-project.org/

Randall, S. M., Ferrante, A. M., Boyd, J. H., & Semmens, J. B. (2013). The effect of data cleaning on record linkage quality. BMC Medical Informatics and Decision Making, 13(1), 64. https://doi.org/10.1186/1472-6947-13-64

Smeets, H. M., de Wit, N. J., & Hoes, A. W. (2011). Routine health insurance data for scientific research: Potential and limitations of the Agis Health Database. Journal of Clinical Epidemiology, 64(4), 424–430. https://doi.org/10.1016/j.jclinepi.2010.04.023

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLoS Medicine, 2(10), e267. https://doi.org/10.1371/journal.pmed.0020267

Wilhelm, J. A., Qiu, M., Paina, L., Colantuoni, E., Mukuru, M., Ssengooba, F., & Bennett, S. (2019). The impact of PEPFAR transition on HIV service delivery at health facilities in Uganda. PLOS ONE, 14(10), e0223426. https://doi.org/10.1371/journal.pone.0223426

World Health Organization. (2007). Everybody's business: Strengthening health systems to improve health outcomes : WHO's framework for action. World Health Organization.

Zheng, Y., Hatakka, M., Sahay, S., & Andersson, A. (2018). Conceptualizing development in information and communication technology for development (ICT4D). Information Technology for Development, 24(1), 1–14. https://doi.org/10.1080/02681102.2017.1396020

International Journal of Community Medicine and Public Health
Singh G et al. Int J Community Med Public Health. 2023 Jan;10(1):148-153
http://www.ijcmph.com
pISSN 2394-6032 | eISSN 2394-6040

DOI: https://dx.doi.org/10.18203/2394-6040.ijcmph20223288

Original Research Article

Development and use of open-source algorithms for space-time emerging hotspot analysis of routine dengue NVBDCP data in Punjab, India

Gurpreet Singh[1]*, Biju Soman[1], Gagandeep Singh Grover[2]

[1]Achutha Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, Kerala, India
[2]Department of Health and Family Welfare, Government of Punjab, India

Received: 18 November 2022
Revised: 01 December 2022
Accepted: 03 December 2022

***Correspondence:**
Dr. Gurpreet Singh,
E-mail: drgurpreet.md.afmc@gmail.com

ABSTRACT

Background: Understanding spatiotemporal epidemiology using open-source and reproducible algorithms add value to routine health information systems. Objectives were to estimate spatial clustering, identify spatial clusters and space-time hotspots of dengue.
Methods: Queen's contiguity neighborhood matrix and row-standardized spatial weights were used. Spatial clustering was estimated using Moran's I. Local Moran's I with sensitivity analysis at 0.01, 0.05, and 0.1 significance levels were performed. The space-time cube model was developed. Gi* statistic and seasonal Mann Kendal test identified persistent and intensifying, persistent, persistent and diminishing, emerging, oscillating, new, historical, and sporadic hotspot sub-districts. Analysis was carried out using R version 4.1.0.
Results: The expected Moran's value was -0.00671. Significant spatial clustering was observed annually in 2016-2018 ($p<0.01$, <0.01, and 0.04, respectively) and was most common in August, followed by July and November. High-high, high-low, low-low, and low-high sub-district clusters were identified between Aug-Dec from 2015-19. Sensitivity analysis highlighted the core and spread of spatial clusters. Faridkot and Muktsar blocks/ sub-districts were persistent and intensifying hotspots.
Conclusions: Spatial clusters were dynamic in space and time. The development of open-source algorithms provides a reproducible and scalable platform for future research and evidence for informed decision-making by public health managers.

Keywords: Spatial correlation analysis, Emerging hotspot analysis, Space-time cube, Routine data, Data science, Dengue

INTRODUCTION

Methods and applications in spatial data analytics have evolved in health sciences in recent decades. They have become an essential tool for epidemiologists for tracking infectious diseases, outbreak analysis, disease surveillance, emergency preparedness and response, environmental health, chronic disease prevention, and community health assessment.[1] These advances provide newer opportunities for public health administrators to plan, analyze, allocate resources, and manage health systems.[2]

Dengue is a viral mosquito-borne disease with more than half of the global population at risk of acquiring the infection through the bite of infected female *Aedes* mosquitoes.[3] It is highly prevalent in tropical and subtropical regions, and the Asian subcontinent

contributes substantially to the global burden. National vector borne disease control programme, India (NVBDCP) reported more than eight lakh dengue cases in the past decade, with a median annual incidence of 6.57 per lakh population. The incidence of dengue was highest in 2019 and 2017 (11.80 and 11.55 per lakh), and the highest median annual incidence of dengue was observed in the state of Punjab (24.49 per lakh). NVBDCP, the nodal programme for controlling vector-borne diseases in the country, maintains statistics on dengue occurrence across states as a part of routine health information systems (RHIS).[4]

There is no specific treatment, and it currently lacks large-scale effective vaccines to prevent and control the increasing dengue burden in India.[3] Understanding the Spatio-temporal patterns of dengue provides insights for estimating patterns of the disease and its association with risk factors in the local context, and thus has the potential for the development of forecasting models for optimizing resource allocation and planning effective vector control measures in low-and middle-income countries. Data science is an interdisciplinary science with utilities like exploratory data analysis, which enables data handling of routine datasets for creating analysis-ready tidy datasets through tools for wrangling, transformation, exploration, etc.[5,6] Further, open-source platforms provide scalable and reproducible algorithms for future use by researchers, epidemiologists, public health managers, and others.

Advancements in open-source geographical information system (GIS) platforms, data handling techniques, and computational resources enable the estimation of spatial clustering and identification of spatial clusters in space and time. Multiple statistical methods have been used for global estimates of spatial clustering depending on the type of data and disease under consideration. Most commonly used methods include Moran's I, Oden's Ipop, Geary's contiguity ratio, and Tango's excess event test for areal spatial data; Edward's K nearest neighbors, Ripley's K function, and Knox test for point spatial data. Similarly, as local indicators to identify spatial clusters, Local Moran's I, Getis-Ord's local Gi, and Gi* statistics for areal data and scan circles (based on Openshaw's, Besag-Newell, Turnbull, and Kulldorf's scan statistics) for point spatial datasets have been used.[7,8] Recently, emerging space-time hotspot analyses have been introduced to understand disease patterns. This looks at the space-time perspective as a modeled cube wherein spatial analysis is carried out for each time slice, and trend analysis of each spatial feature is carried out over time.[9] However, earlier studies have used proprietary software, and to the best of our knowledge, there is a lack of open-source algorithms for the same.

Therefore, the present study was carried out to estimate the spatial clustering of dengue in the state of Punjab, identify spatial clusters (sub-districts), develop an open-source algorithm for emerging space-time hotspot analysis, and provide empirical evidence using the dengue RHIS dataset.

METHODS

The study design of the present study was ecological study with a data science approach. The study included spatial autocorrelation analysis and space-time emerging hotspot analysis of secondary data. The anonymized line list of lab-confirmed dengue cases reported by NVBDCP, Punjab, from 2015-19, was analyzed. The line listing data included cleaned, standardized, and geocoded addresses. Point in polygon analysis was carried out, and population projections based on census 2011 were calculated to estimate monthly dengue incidence rates at the sub-district level. The spatial file of the sub-district (Block) multi-polygons was obtained from Punjab remote sensing authority. The inclusion criteria for the present study were lab-confirmed dengue cases reported by directorate health services. Those records where geocoded addresses and testing dates missing after pre-processing of raw data were excluded from the study. Total number of reported lab-confirmed cases during the study period was 64,454. After excluding cases with no location/ time details, 63,741 cases (98.8%) were included in study for analysis.

Spatial autocorrelation analysis

The neighborhood matrix was constructed based on areal units with contiguous boundaries. The neighborhood was defined according to the queen's contiguity wherein the sub-districts were neighbors when they had at least one shared boundary point. The row-standardized spatial weights were calculated. Annual and monthly spatial autocorrelation analyses were carried out. Moran's I statistic provided global estimates for spatial clustering, and Local Moran's I was calculated to identify spatial clusters. Sub-district categorized in LISA quadrants of high-high, high-low, low-low, low-high, and unclassified based on the dengue incidence in the sub-district under consideration and its neighbors as compared to the global estimates. Sensitivity analyses at 0.01, 0.05, and 0.1 levels of significance were carried out to determine the core and spread of spatial clusters at a given point in time.

Space-time emerging hotspot analysis

A space-time cube model was constructed wherein the space was defined as areal units (sub-districts) and time as monthly intervals. For each sub-district, a monthly time series was constructed, and trend analysis was performed using the seasonal Mann-Kendall trend test. A trend was said to be significantly positive when the z score was positive and the $p \leq 0.05$. For each month in the monthly time series, Getis Ord Gi* statistic was calculated for all sub-districts. The sub-districts were considered as hotspot/ coldspot for the respective month when the calculated z score was $\geq 1.96/\leq 1.96$, respectively. Based on the space-time patterns, sub-districts were categorized into eight categories.

The definitions for space-time classification were adapted from ArcGIS literature for emerging hotspot analysis.[10] A sub-district was categorized as a 'persistent' hotspot when it has been a hotspot for at least a month for four years between 2015-19, 'persistent and intensifying'/'persistent and diminishing' when a significant positive/ negative time trend was present in addition to above respectively. Sub-districts that were hotspots only in recent years (at least twice in 2017-19) were categorized as 'emerging,' those that were hotspots earlier but have ceased to be in recent years as 'historical', hotspots in 2019 but never earlier as 'new', hotspot only once as 'sporadic', a hotspot in on-off fashion as 'oscillating' and others as 'not categorized'.

Permissions and clearances

Present study was a part of a more extensive study being undertaken as a Ph.D. project by the first author. Ethical approval was obtained from institutional ethics committee (IEC/IEC-1653; IEC reg No. ECR/189/Inst/KL/2013/RR-16), and study has been registered on clinical trials registry of India (CTRI/2021/01/030245). Permission from the directorate of health services, Punjab and Punjab remote sensing Authority obtained. Detailed study protocol and algorithms for pre-processing datasets using reproducible open-source algorithms have been published elsewhere.[11-13]

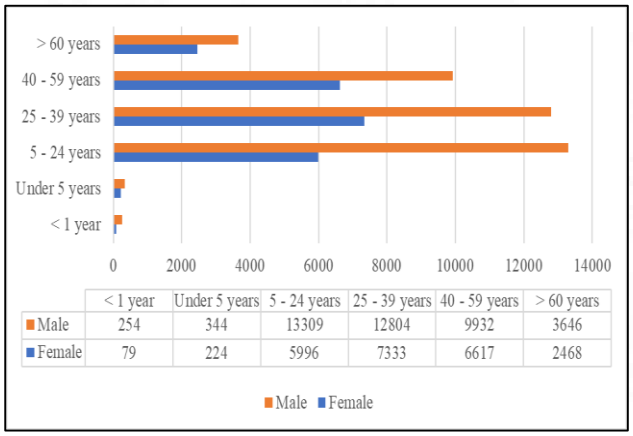
Statistical software

The analysis was carried out using R version 4.1.0 and included the use of tidyverse, lubridate, sf, spdep, rgeoda, and Kendall packages.

RESULTS

Age and sex distribution of reported cases

The age and sex distribution of reported cases are presented as Figure 1. Most cases were males (64%) and in the age group of 25-39 years (32%).



	< 1 year	Under 5 years	5 - 24 years	25 - 39 years	40 - 59 years	> 60 years
Male	254	344	13309	12804	9932	3646
Female	79	224	5996	7333	6617	2468

Figure 1: Age and sex distribution of reported dengue cases.

Neighborhood matrix

The neighborhood matrix features of the sub-district spatial file are represented in Table 1. The matrix was symmetrical without isolated areal units. There were 150 areal units, with the majority (37) sub-districts having five neighbors. The number of links varied from one to nine, and the mean number of links was 5.28.

Table 1: Neighborhood matrix features.

Neighborhood matrix features	Characteristics
Number of regions (sub-districts)	150
Number of nonzero links	792
Percentage nonzero weights	3.52
Average number of links	5.28
Minimum number of links	01
Sub-districts with minimum number of links	02 (Bamial and Sardulgarh)
Maximum number of links	09
Sub-districts with maximum number of links	02 (Ludhiana-I and Batala)

Spatial clustering of dengue

Global Moran's I was found to be statistically significant at multiple occasions during the study period. The space-time heatmap for the presence of spatial clustering is represented in Figure 2. It was observed that the spatial clustering was predominantly during the seasonal onset and waning periods for dengue occurrence in July-Aug and November, respectively. The duration of significant spatial clustering for a given annual timestamp was dynamic and varied across the years. The shortest duration of spatial clustering of dengue incidence was observed in 2019 (July-August), followed by the year 2015 (June-August and November) and 2016 (August-November). An on-off pattern of spatial clustering was observed in the year 2018.

Spatial clusters of dengue

LISA statistics provided evidence for the dynamic nature of both spatial clusters (high-high and low-low) and spatial outliers (high-low and low-high) across time. The sensitivity analysis enabled the identification of the core and spread of the clusters. The sensitivity analysis results at 0.01, 0.05, and 0.1 level of significance as a local significance map are represented in Figure 3. The patterns showed the highest density of high-high (Sirhind, Majri, and Sanaur), low-low (Batala, Ajnala, and Shahkot), low-high (Shambu Kalan, Khera, and Morinda), and high-low (Taran-Taran, Bathinda, and Verka) spatial clusters for dengue occurrence across sub-districts.

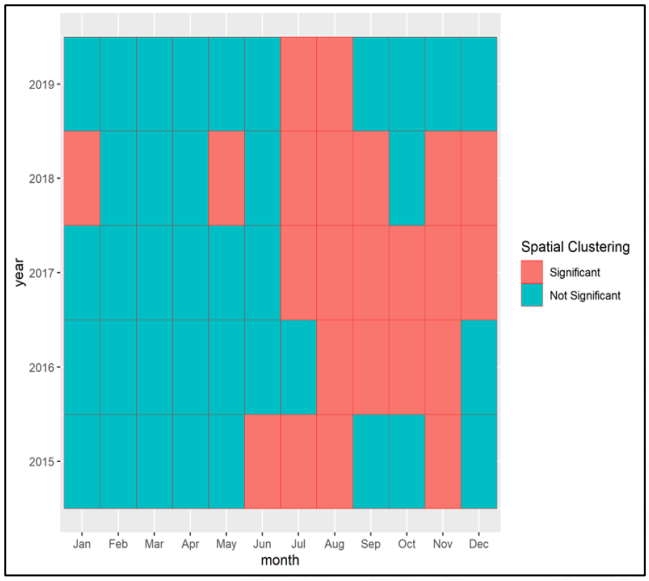


Figure 2: Space-time heatmap of global estimates for spatial clustering (Moran's I).

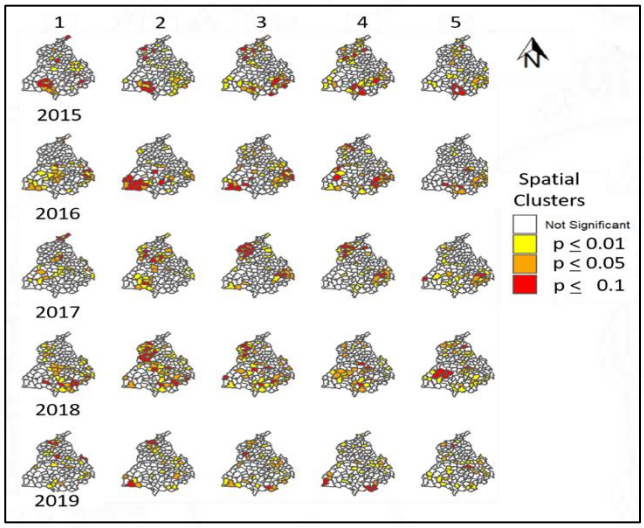


Figure 3: Local significance map for sensitivity analysis (Local Moran's I).

Space-time emerging hotspot analysis

Findings from space-time emerging hotspot analysis at monthly intervals are represented in Figure 4. It was observed that the numbers of hotspots were more persistent in the southwestern and south-eastern region; however, the sub-districts on the western border showed the presence of hotspots in recent years during the study period.

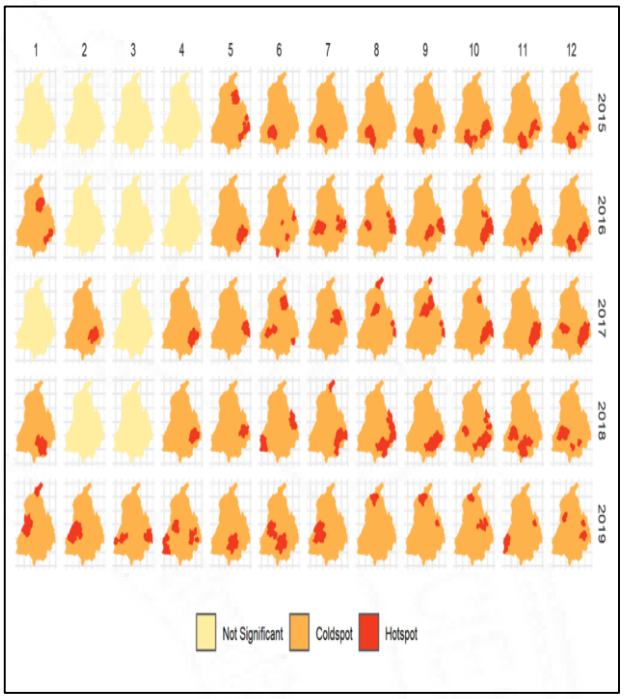


Figure 4: Space-time hotspot analysis grid (Gi* statistic).

The details of the identified hotspots are represented in Table 2. A mixed pattern of sub-districts belonging to all categories of emerging hotspot analyses were found to be existing in the state. The majority (n=27) of sub-districts were sporadic hotspots, followed by persistent hotspots (n=21), and new hotspots (n=21).

Table 2: Classification of sub-districts.

Hotspot category	Frequency, n=150 (%)	Sub-districts
Persistent and intensifying	02 (1.33)	Faridkot and Muktsar
Persistent	21 (14.00)	Amloh, Bagha Purana, Balachaur, Bassi Pathanan, Bhawanigarh, Bhikhi, Derabassi, Dhuri, Jaitu, Kharar, Khera, Kot Kapura, Majri, Morinda, Nabha, Patiala, Rajpura, Rampura, Samana, Shambu Kalan, and Sirhind
Persistent and diminishing	02 (1.33)	Ghanaur and Sanaur
Emerging	14 (9.33)	Abohar, Arniwal, Aur, Dhar Kalan, Fazilka, Gharota, Khamanon, Khuian Sarwar, Ludhiana II, Narot Jaimal Singh, Pathankot, Sangrur, Saroya, and Sujanpur
Oscillating	10 (6.67)	Adampur, Barnala, Bhogpur, Budhlada, Hoshiarpur I, Hoshiarpur II, Jhunir, Machhiwara, Mansa, and Maur

Continued.

Hotspot category	Frequency, n=150 (%)	Sub-districts
New	21 (14.00)	Batala, Dera Baba Nanak, Dhariwal, Dina Nagar, Fatehgarh Churian, Firozpur, Ghall Khurd, Gurdaspur, Guru Har Sahai, Jalalabad, Kalanaur, Khanna, Mahal Kalan, Makhu, Mamdot, Moga II, Naushera Pannuan, Patti, Sehna, Valtoha, and Zira
Historical	06 (4.00)	Bhunga, Jalandhar East, Jalandhar West, Malerkotla, Nawan Shahr, and Tanda
Sporadic	27 (18.00)	Ahemdagarh, Anandpur Sahib, Andana, Banga, Bathinda, Bhunarheri, Chamkaur Sahib, Dhilwan, Dirba, Garh Shankar, Goiniana, Kapurthala, Kot Bhai, Gidderbaha, Lehra Gaga, Nakodar, Nathana, Nurpur Bedi, Patran, Phagwara, Rupnagar, Sangat, Sardulgarh, Sher Pur, Sudhar, Sultanpur Lodhi, Sunam, and Talwandi Sabo

DISCUSSION

The present study documents evidence of the potential of routine health data to understand spatiotemporal patterns and the development of open-source algorithms for space-time emerging hotspot analysis using dengue line listing from NVBDCP, Punjab, as empirical datasets. Integrating data science approaches in routine health information systems is expected to help health authorities to enhance dengue preventive strategies and develop public health interventions targeted for the identified cluster areas.[2,14]

Adoption of geographic information systems accelerated rapidly in United States after the 1970s due to availability of open data-sharing platforms.[1] The launch of the open data initiative by the government of India, WHO-IDSP recommendations on increased use of GIS platforms, national digital health mission and other related initiatives in India provide opportunities for researchers and administrators to collaborate for understanding disease epidemiology and to incorporate advances in Spatio-temporal epidemiology, for resource allocation.[15-17]

The emergence of high-low spatial outliers at the seasonal onset was seen in predominantly urban areas surrounded by rural sub-districts, followed by the subsequent spread of high dengue incidence. Also, the dynamic shift of high dengue incidence across sub-districts was observed over time. In a study carried out in Delhi, India, Olivier et al described and compared a similar phenomenon to a 'forest fire' signature wherein the spread of dengue occurs rapidly and a local cluster before being burned out seed adjacent areas, even at a location beyond the flight range for the mosquito dispersal.[18] Similar findings in the present study strengthen evidence for institutionalization of GIS into health care to develop decision support systems. Also, such findings urge future research work incorporating non-health sector routine data for understanding nuances of associated environmental, climatic, socio-demographic and health system factors for risk-based resource allocation in health care which is a hierarchical system with multiple decision makers.[19]

It is essential to understand that the selection of the operational definition for creating the neighborhood matrix should be based on the transmission patterns and epidemiology of the disease under investigation.[8] The present study created a neighborhood matrix using Queen's contiguity. Dengue, being a mosquito-borne disease, and considering the mobility patterns of the population, the areal units sharing even a single boundary point should be considered.[7] This is in contrast to the approaches for modeling in studies in the domain of one-health. Compared to the disease transmission dynamics in plants, zoonoses, and other inter-sectoral areas, human mobility with increased connectivity and commutations for work and leisure, the selection of neighborhood matrices needs deliberate caution and consideration.[7,8]

The representation of the findings as a static figure provides a limited interpretation of space-time dynamic disease processes. Technological advancements need to be harnessed and applied in public health to become appropriate, affordable, and acceptable for benefits to the weaker sections of society for sustainable development. The development of interactive automated parameterized dashboards enables better understanding and fosters evidence-informed decision-making.[17] For the same, intersectoral collaborations between health care and information technology professionals is the need of the hour, evident across multiple sectors, including health care, and increasing exponentially.[17,18]

The study modeled data based on routinely reported cases of lab-confirmed dengue cases in the government setup. Sub-clinical and mild clinical manifestations are seen in up to 80% of dengue infections, and the same could not be captured in the present study.[3] However, the assumption of randomness for varied clinical patterns across sub-districts justifies estimating disease patterns using lab-confirmed line listing data. Future research on clinical patterns using serological studies is required and beyond the scope of the present study. Also, point pattern analysis could not be carried out in the study undertaken. This may be attributed to the lack of coordinates of case occurrence in the RHIS. The addresses were geocoded; however, the bounding boxes obtained for the geocoded data though found adequate for areal data analysis, and point pattern analysis was not recommended. Standardization of address components in RHIS is required to allow such inferences in future studies. Also, similar to the recent introduction of the integrated health

information portal by GoI, GIS integration into RHIS shall enable point pattern analysis in future studies based on the data science approach.

CONCLUSION

Spatial clustering was observed at the extremes of the seasonality pattern of dengue incidence. Spatial clusters were dynamic in space and time. The development of open-source algorithms provides evidence for informed decision-making by public health managers. Research on routinely collected data has the potential to provide insights into the data quality issues, spatio-temporal disease epidemiology, and identification of features/ variables for disease modeling in subsequent studies.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: The study was approved by the Institutional Ethics Committee

REFERENCES

1. Davenhall WF, Kinabrew C. Geographic Information Systems in Health and Human Services. In: Kresse W, Danko D, eds. Springer Handbook of Geographic Information. Cham: Springer International Publishing. 2022;781-805.
2. Garg PK. Geospatial Health Data Analytics for Society 5.0. In: Garg PK, Tripathi NK, Kappas M, Gaur L, eds. Geospatial Data Science in Healthcare for Society 5.0. Singapore: Springer Singapore. 2022;29-58.
3. World Health Organization. Dengue and severe dengue Factsheet. Available at: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue. Accessed on 27 Sept, 2022.
4. Directorate of National Vector Borne Disease Control Programme. Long Term Action Plan for prevention and control of Dengue and Chikungunya. 2007. Available at: https://nvbdcp.gov.in/Doc/Final_long_term_Action_Plan%20.pdf . Accessed on 27 Sept, 2022.
5. Wickham H, Grolemund G. R for data science: import, tidy, transform, visualize, and model data, First edition. Sebastopol, CA: O'Reilly. 2016.
6. Van der Aalst W. Data Science in Action. In: van der Aalst W, ed. Process Mining: Data Science in Action. Berlin, Heidelberg: Springer. 2016;3-23.
7. Pfeiffer DU, Robinson TP, Stevenson M, Stevens KB, Rogers DJ, Clements ACA. Spatial Analysis in Epidemiology. OUP Oxford. 2008.
8. O'Sullivan D, Unwin D. Geographic Information Analysis. Wiley. 2014.
9. Emerging Hot Spot Analysis: Finding Patterns over Space and Time. Azavea. 2017;15.
10. ArcGIS Pro. Emerging Hot Spot Analysis (Space Time Pattern Mining). Available at: https://pro.arcgis.com/en/pro-app/latest/tool-reference/space-time-pattern-mining/emerginghotspots.html. Accessed on 7 Oct, 2022.
11. Singh G, Mitra A, Soman B. Development and use of a reproducible framework for spatiotemporal climatic risk assessment and its association with decadal trend of dengue in India. Indian J Community Med. 2022;47:50.
12. Singh G, Soman B. Spatiotemporal Epidemiology and Forecasting of Dengue in the state of Punjab, India: Study Protocol. Spatial Spatio-temporal Epidemiol. 2021.
13. Singh G, Soman B, Mitra A. A Systematic Approach to Cleaning Routine Health Surveillance Datasets: An Illustration Using National Vector Borne Disease Control Programme Data of Punjab, India. arXiv:210809963 [cs] 2021;23.
14. Hung YW, Hoxha K, Irwin BR, Law MR, Grépin KA. Using routine health information data for research in low- and middle-income countries: a systematic review. BMC Health Services Res. 2020;20:790.
15. Open Government Data (OGD) Platform India. Open Government Data (OGD) Platform India. 2022. Available at: https://data.gov.in. Accessed on 29 Sept, 2022.
16. Directorate General of Health Services, India. Joint Monitoring Mission Rep. 2015.
17. National Digital Health Mission. Available at: https://ndhm.gov.in/. Accessed on 19 Feb, 2021.
18. Telle O, Vaguet A, Yadav NK. The Spread of Dengue in an Endemic Urban Milieu-The Case of Delhi, India. PLoS one. 2016;11:e0146539.
19. Yan Z, Haimes YY. Risk-based multiobjective resource allocation in hierarchical systems with multiple decisionmakers. Part I: Theory and methodology. Syst Engin. 2011;14:1-16.
20. Yigitbasioglu OM, Velcu O. A review of dashboards in performance management: Implications for design and research. Int J Accounting Information Systems. 2012;13:41-59.
21. Few S. Information dashboard design: the effective visual communication of data, 1st ed. Beijing; Cambride [MA]: O'Reilly, 2006.

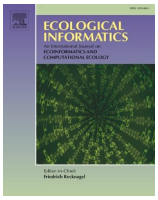
Cite this article as: Singh G, Soman B, Grover GS. Development and use of open-source algorithms for space-time emerging hotspot analysis of routine dengue NVBDCP data in Punjab, India. Int J Community Med Public Health 2023;10:148-53.

Ecological Informatics 75 (2023) 102020

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Exploratory Spatio-Temporal Data Analysis (ESTDA) of Dengue and its association with climatic, environmental, and sociodemographic factors in Punjab, India

Gurpreet Singh [a], Biju Soman [a,*], Gagandeep Singh Grover [b]

[a] Achutha Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, Kerala, India
[b] Department of Health and Family Welfare, Government of Punjab, India

ARTICLE INFO

Keywords:
Exploratory data analysis
Exploratory Spatio-Temporal Data Analysis
Dengue
Spatial analysis
Time series analysis
Reproducible research

ABSTRACT

Routine health data is rich in information but underutilised for research in Low and Middle-Income countries. The present study was carried out to understand spatiotemporal patterns of dengue and its association with risk factors using routine data from health and allied sectors. ESTDA included estimation of dengue incidence rates, time series features and correlation coefficients up to the sub-district level. Scatter plots and correlation coefficients were used to identify relationships between covariates. Dengue incidence in 2015–19 was 47.76, 33.64, 52.03, 49.71, and 33.36 per 100,000, respectively, with a mean (SD) age of 34.33 (16.78) years and the majority being males (63.94%). Dengue had significant cross-correlation, non-linear relationships, and spatio-temporal associations with climatic, environmental, and socio-demographic risk factors. Significant autocorrelation of dengue occurrence was present at a lag of one month with seasonal patterns. The reproducible open-source algorithms add value to existing Routine Health Information Systems, and the findings will enable the development of Spatio-temporal models in future research. The research was done using R version 4.1.0.

1. Introduction

"Systems that comprise data collected at regular intervals at public, private, and community-level health facilities and institutions and health programs" are Routine Health Information Systems (RHIS). (MEASURE Evaluation, 2022) RHIS systematically collects data, often for administrative use within health departments. Despite being rich in information, RHIS is underutilised for research in low- and middle-income countries, especially in the Asian subcontinent.(Hung et al., 2020).

The use of interdisciplinary approaches by linking data from non-health sectors and GIS is recommended by WHO-India Joint Monitoring Mission to strengthen disease surveillance.(Directorate General of Health Services, India, 2015) Data science is "an interdisciplinary field involving processes, theories, concepts, tools, and technologies, that enable the review, analysis, and extraction of valuable knowledge and information from structured and unstructured (raw) data.(Data Science - MeSH - NCBI, 2023) Thus, data science allows researchers to explore and analyse routine health data and its associations with ecological datasets. (van der Aalst, 2016).

Dengue is the fastest growing Vector Borne Disease (VBD) globally, and the highest burden is present in Low- and Middle- Income Countries (LMICs). South-East Asian Region (SEAR) contributes to 52% of the global dengue burden and is a significant public health problem in India. Dengue incidence in a population is determined by climatic, environmental, and socio-demographic conditions. These associations vary from place to place, and it is essential to understand them in the local context to develop Spatio-temporal models for evidence-informed public health decision-making.

Exploratory Data Analysis (EDA) is a comparatively new area of statistics.(Bruce and Bruce, 2017) It is based on the principle that "It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it"(Tukey, 1977). EDA is complementary to confirmatory inferential statistics as it minimises violations of assumptions for model building. It also enables understanding of the data and guides appropriate questions, analysis, and models. Exploratory Spatial Data Analysis (ESDA) is an advancement to EDA for the detection of spatial patterns, hypotheses formulation based on spatial features, and assessing the appropriate spatial models. Similarly, for datasets including both space and time attributes, Exploratory Spatio-

* Corresponding author.
E-mail address: bijusoman@sctimst.ac.in (B. Soman).

https://doi.org/10.1016/j.ecoinf.2023.102020
Received 31 July 2022; Received in revised form 4 February 2023; Accepted 4 February 2023
Available online 10 February 2023

Temporal or Space-Time Data Analysis (ESTDA) has been recently introduced and is an active research domain in the field of Geographic Information Science (GIS). (De Smith et al., 2018) The current approach is preferred compared to other methods for time series classification, such as federated distillation learning system (EFDLS)(Xing et al., 2022a), robust temporal feature network (RTFN)(Xiao et al., 2021), and strategies for hybridisation of supervised learning, unsupervised learning, and Self distillation(Xing et al., 2022b). The mathematical foundations for the approach used are underpinned by the fact that dengue counts in routine data have Poison distribution and have been explored to subsequently develop spatiotemporal regression models and time series forecasting for a continuous outcome/ dependent variable. Further, underreporting of mild and missed cases in LMICs and the lack of adequate geocoding accuracy for spatial point pattern analysis pose a limitation for classification algorithms.

Increasing initiatives for improving RHIS data quality in LMICs have been undertaken in the past few decades; however, neglected tropical diseases such as dengue require additional efforts to understand the data structure, quality issues, and mechanisms by which evidence can be generated from the routinely collected data.

India contributes to around 34% of the global burden of dengue. (Dengue and severe dengue, 2023) However, the association of dengue with risk factors in the country is poorly explored. Also, because of multiple climatic zones and varied ecology and socio-demography within the states in India, the exploration of such associations in local context is recommended.(Kakarla et al., 2019) Further, in the majority of studies carried out in the country on dengue and its risk factors using RHIS, the association of dengue has been explored for associations with either a single risk factor, or at a spatial granularity of national or state level or with isolated spatial or time series analysis. There is an unmet need for studies exploring dengue association with multiple climatic, environmental, and socio-demographic factors at higher resolution for understanding disease dynamics. There is an evident lack of studies to understand dengue epidemiology using satellite imagery datasets, its space-time associations with multiple risk factors, and the use of advancements in technology which can provide decision support to public health managers. There is also a dearth of open-source solutions to foster research using RHIS. Additionally, within the health sector, on the one hand, there are robust RHIS such as Nickshay for Tuberculosis, digitisation for Reproductive and Maternal Health, etc., on the other, routine data for Dengue occurrence still exists in excel sheets is entered manually, and poses challenges to its use in research. Therefore, the present study was thus undertaken as ESTDA in data science using dengue routine health data and open-source software to explore patterns and associations for dengue incidence and its climatic, environmental, and socio-demographic risk factors in Punjab, India.

2. Material and methods

The present study included secondary data analysis of routinely collected datasets by healthcare system and multiple open-source datasets. The study design was an ecological study in healthcare epidemiology using the data science approach.

2.1. Data sources/ variables

Pre-processed and cleaned routine lab-confirmed dengue line listing data of NVBDCP, Punjab, from 2015 to 19 was used to explore the spatiotemporal epidemiology of Dengue up to the sub-district level. Climatic and environmental data analysis was conducted using pre-processed satellite imagery datasets. Analysis of socio-demographic factors was based on Census 2011 and spatial datasets provided by Socioeconomic Data and Applications Centre (SEDAC), Columbia University. Sub-district level spatial dataset was obtained from Punjab Remote Sensing Authority. The dengue occurrence was considered a dependent variable, and the climatic, environmental, and socio-demographic factors as independent variables.

2.2. Statistical analysis

ESTDA included the estimation of annual, quarterly, monthly, and weekly dengue incidence rates for sub-subdistricts. The noise in the time series incidence rates was estimated using Hurst coefficient and spectral entropy measures. The extent of lag associations for dengue occurrence was explored using Autocorrelation analysis and the extent of seasonality and trend in dengue was estimated using Time-series decomposition analysis. Risk mapping of dengue in the state was done using Standardised Incidence Ratios. Age and gender distribution of dengue occurrence were estimated.

Similarly, ESTDA of climatic and environmental risk factors and exploratory spatial analysis of purely spatial features were carried out. Multiple statistical measures were calculated to understand the association of dengue with risk factors. Pearson's correlation coefficient, time series cross-correlation function, and visualisations using scatter plots and space-time Hovemoller diagram were carried out.

2.3. Statistical software

The analysis was carried out using R version 4.1.0 and included the use of tidyverse(Wickham et al., 2019), sf(Pebesma, 2018), spdep (Bivand et al., 2013), and fpp3(Forecasting: Principles and Practice (3rd Ed), 2021) packages.

2.4. Ethics and permissions

The study was a part of a larger study being undertaken as a Ph.D. project by the first author. Ethical approval was obtained from Institutional Ethics Committee (IEC/IEC-1653; IEC Reg No. ECR/189/Inst/KL/2013/RR-16), and the study has been registered on the Clinical Trials Registry of India (CTRI/2021/01/030245). Permission from the Directorate of Health Services, Punjab, Punjab Remote Sensing Authority, National Remote Sensing Centre, India, and Earth data, NASA, were obtained. A detailed study protocol and algorithms for pre-processing routine dengue datasets using reproducible open-source algorithms have been published elsewhere.(Singh et al., 2021, 2022).

3. Results

3.1. Dengue epidemiology in the state

The state dengue incidence rates in 2015–19 were 47.8, 33.6, 52.0, 49.7, and 33.4 per 100,000, respectively. The mean (SD) age of the reported dengue cases was 34.3 (16.8) years. Most patients were males (63.9%) and in the age group of 25–39 years (32.0%). The fourth quarter, October, and week 41 to week 46 were dengue's peak periods. The Hurst coefficient of quarterly, monthly, and weekly time series was 0.5, 0.91, and 0.99, and the spectral entropy measure was 0.28, 0.53, and 0.72, respectively. Time series ACF showed significant positive autocorrelation at the lag of four quarters, one month, and up to 6 weeks. The PACF showed significant positive autocorrelation of the residuals at one-month and one-week lags. The annual dengue incidence across districts varied from 4.75 per lakh (Fazilka district in 2016) to 210 per 100,000 (Sahibzada Ajit Singh Nagar district in 2017). Space-time distribution of dengue incidence at monthly intervals across sub-districts is represented in Fig. 1. As illustrated in the figure, the seasonality of dengue incidence is evident; however, the northern sub-districts have shown rising incidence in recent years. On disease risk mapping, SIR $\geq$1.5 for $\geq$10 months was found in Sangrur, Sirhind, Mansa, Bassi Pathanan, Nawan Shahr, Patiala, Kotkapura, Aur, and Rupnagar sub-districts.

The findings of the time series STL decomposition analysis are represented in Table 1. Dengue incidence, LST (Day), LST (Night),

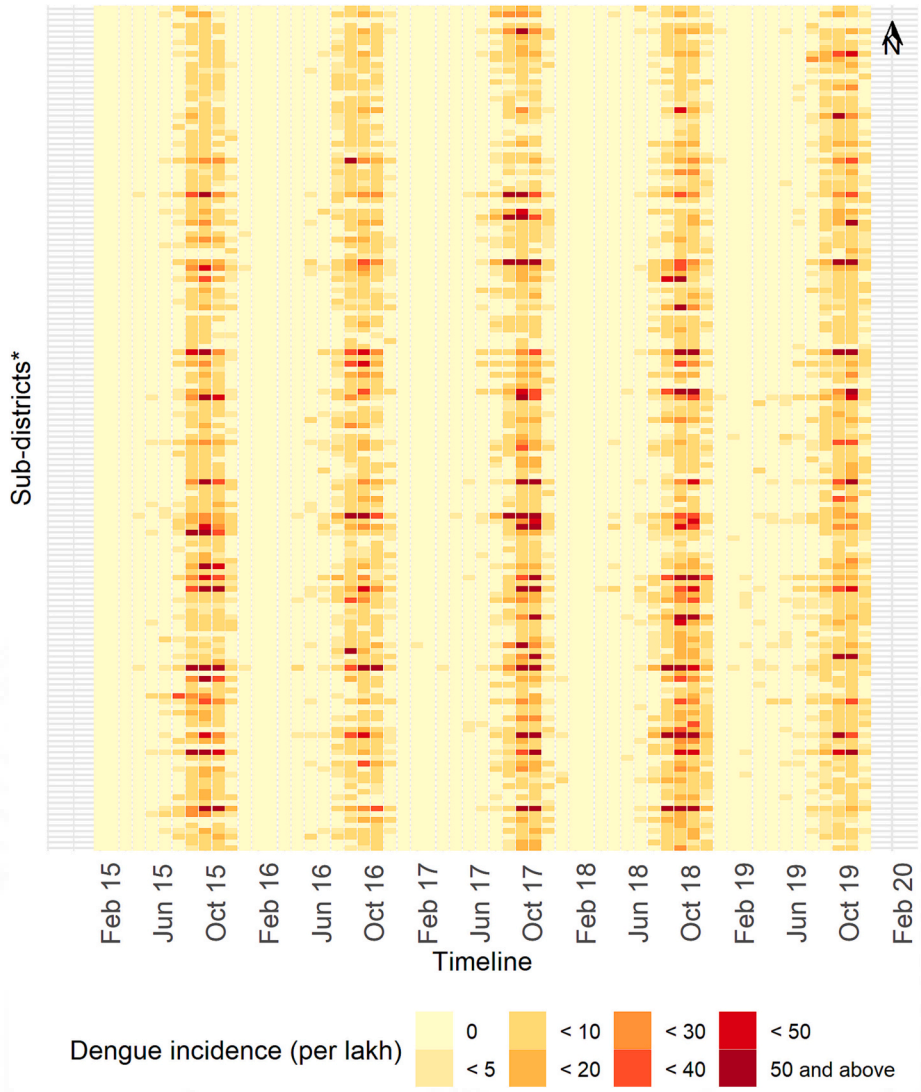


Fig. 1. Space time distribution of dengue incidence across sub-districts.
The X axis depicts timeline; Y axis represents spatial features; and the color represents the dengue incidence at a given time and place.
** The 150 subdistricts in the state are arranged from North to South on the Y axis from top to bottom in the figure.*

Table 1
Findings of Time series STL decomposition.

Time series features	Seasonality strength	Trend strength
Dengue incidence		
Quarterly	0.93	0.31
Monthly	0.91	0.14
Weekly	0.85	0.07
Land Surface temperature		
Night	0.99	0.32
Day	0.98	0.20
Precipitation	0.80	0.11
Relative Humidity	0.80	0.40
NDVI	0.87	0.28

precipitation, relative humidity, and NDVI values showed strong seasonality and weak trends in the state.

Fig. 2 represents the distribution of time-averaged climatic factors across sub-districts. The southwestern and southeastern sub-districts had higher mean temperatures (day and night). Further, the southwestern sub-districts had lower cumulative rainfall per sq. km. Higher mean relative humidity was observed in eastern subdistricts, and higher mean NDVI values were observed in southeastern and northern subdistricts.

The characteristics of the spatial features are represented in Table 2. Elevation levels were highest in the northeastern bordering sub-districts, and the slope was highest in the northeast region and declined towards the southern part of the state. The urban pockets and built-up area distribution was similar, with high density in the southern and southeastern region. The percentage of spatial grid cells with a built-up area at a 50% threshold in a sub-district varied from zero to 32.5%. The lowest built-up area was in Bhunarheri, Ghanaur, Talwara, and Dhar Kalan subdistricts and the highest was in Ludhiana-I (32.5%) followed by Verka (23.8%), and Jalandhar East (14.13%).

3.2. Correlation between risk factors

The correlation matrix between land surface temperatures and additional climatic risk factors is represented in Fig. 3. There was a significant correlation between LST (night) with LST (day), the maximum and minimum temperature at two meters, WET bulb temperature, the maximum and minimum temperature at ten meters, and dew point temperature (at r values of 0.80, 0.91, 0.98, 0.90, 0.97, 0.91,

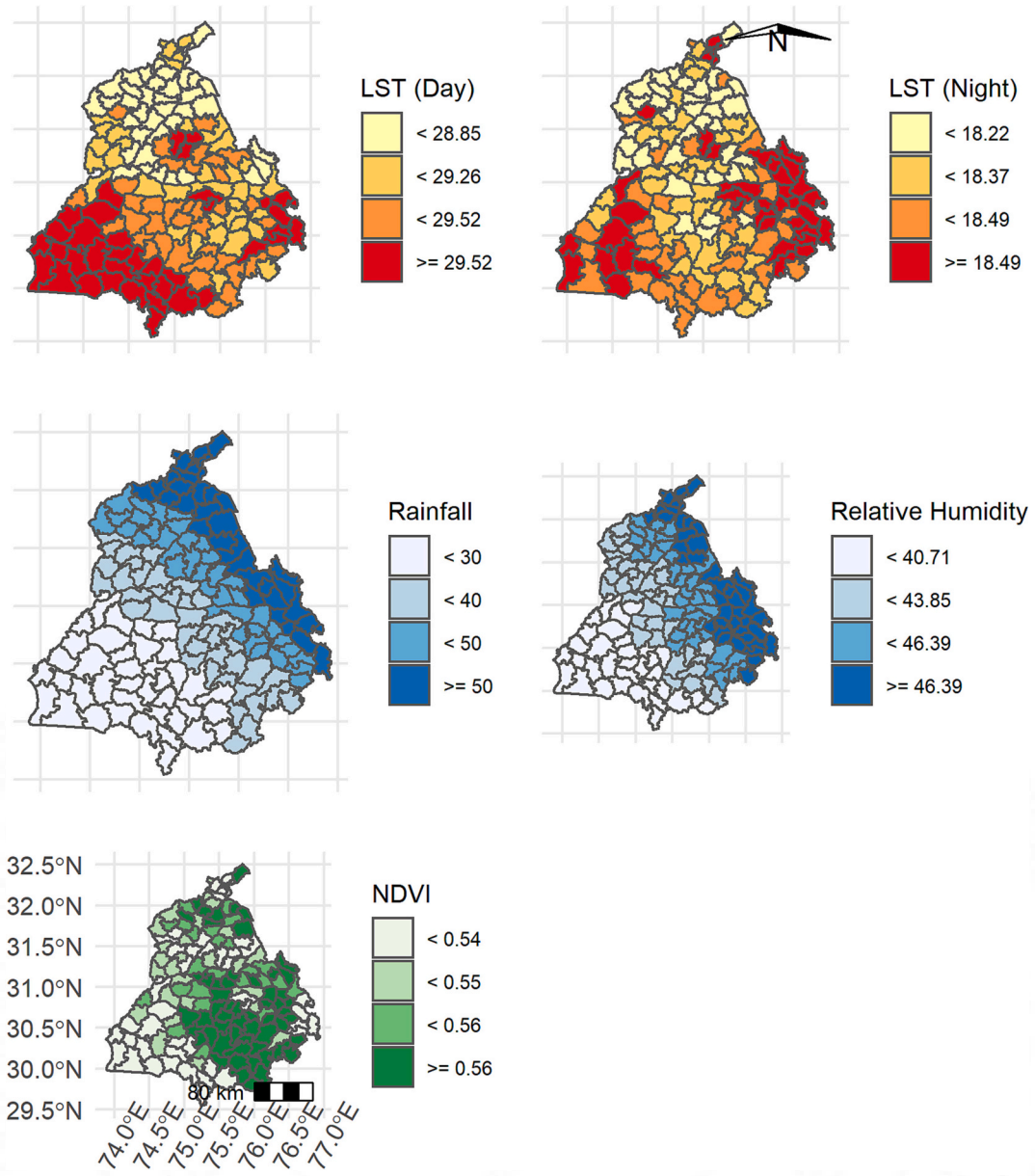


Fig. 2. Spatial distribution of time averaged features across sub-districts.

0.90, respectively, $p < 0.05$). Also, the correlation between elevation and slope, urbanisation and built-up area, household density, and the female literacy rate was statistically significant ($r = 0.76$, 0.8, 0.43, $p < 0.05$).

3.3. Association between dengue and its risk factors

The time series cross-correlation coefficients and scatter plots at specified lags for land surface temperature (night), cumulative precipitation, relative humidity, and NDVI with dengue incidence are represented in Fig. 4. Positive and significant cross-correlations were present for land surface temperature (night), cumulative precipitation, relative humidity, and NDVI with dengue incidence from the lag of 2 to 5 months ($r = 0.52$, 0.60, 0.59, and 0.45), 2 to 4 months ($r = 0.60$, 0.76, and 0.44), 1 to 3 months ($r = 0.35$, 0.68, and 0.56), and at one month ($r = 0.25$), respectively ($p < 0.05$). Further, the scatter plots suggested non-linear relationships between these variables and dengue incidence. Also, the correlation of urbanisation, built-up area, population density, and persons per household was significant with dengue incidence ($r = 0.34$, 0.28, 0.23, and -0.17, respectively, $p < 0.01$).

4. Discussion

Dengue occurrence in a population is influenced by ecological and socio-demographic factors (Farrar and Manson, 2014). Exploration of associations between disease occurrence and its eco-socio-demographic factors provides evidence for feature selection and development of early warning forecasting systems for strengthening disease surveillance, efficient resource allocation, and timely implementation of prevention and control measures. The present study is a first-of-kind study from the state which documents the potential of using RHIS from empirical dengue data maintained by the health directorate and its data linkages with satellite imagery (ecological factors) and census data (socio-demographic factors) for understanding dengue epidemiology. Further, the use of open-source platforms and reproducible algorithms in the present study are aligned with the open data policy provided by World Health Organization and National Data Policy and thus provide scalable algorithms for future use in LMIC settings and data analytics in

Table 2
Characteristics of spatial features.

Spatial features	Characteristics
Environmental factors	
Elevation (meters)	
Mean	198.8
Range	132.7–513.9
Sub-districts with the lowest values	Fazilka (132.6), Khuian Sarwar (136.3), Jalalabad (138.9)
Sub-districts with the highest values	Dhar kalan (513.9), Talwara (433.6), Bhunga (322.3)
Slope (degrees)	
Mean	2.6
Range	1.5–10.6
Sub-districts with the lowest values	Bhunarheri (1.5), Khera (1.6), Ghanaur (1.7)
Sub-districts with the highest values	Dhar Kalan (10.6), Talwara (8.1), Hoshiarpur II (6.1)
Socio-demographic factors	
Level of urbanisation (%)	
Mean	6.1
Range	0–56.7
Sub-districts without urban areas	16 (10.67%)
Sub-districts with the highest values	Verka (56.7), Ludhiana-I (45.5), Jalandhar -East (25.5).
Built up area at 50% threshold (%)	
Mean	
Range	
Sub-districts with the lowest values	
Sub-districts with the highest values	
Population density (per sq km)	
Mean	599
Range	248–4959
Sub-districts with the lowest values	Dhar Kalan (248), Lambi (269), Sangat (273)
Sub-districts with the highest values	Verka (4959), Ludhiana-I (4751), Jalandhar east (2532)
Household density (per sq km)	
Mean	74
Range	42–175
Sub-districts with lowest values	Makhu (42), Muktsar (43), and Sultanpur Lodhi (44)
Sub-districts with the highest values	Pathankot (175), Gharota (144), and Sujanpur (139)
Female literacy rate	
Mean	59.4
Range	43.9–74.9
Sub-districts with the lowest values	Lehra Ganga (43.9), Valtoha (44.1), Jhunir (44.6)
Sub-districts with the highest values	Talwara (74.9), Bhunga (73.5), Hajipur (72.8)
Person per household	
Mean	5.1
Range	4.5–5.7
Sub-districts with the lowest values	Adampur (4.5), Tanda (4.5), Talwara (4.6)
Sub-districts with the highest values	Valtoha (5.7), Gandiwind Tatla (5.7), Andana (5.6)

developing routine data-driven models (National Data Sharing and Accessibility Policy | Department Of Science and Technology, 2022).

Routine Health Information Systems capture the place and time of occurrence of a health event/ disease (Jamison and World Bank and Disease Control Priorities Project, 2006). This enables the use of Spatio-temporal methods in understanding disease epidemiology. In the present study, high annual dengue incidence (33.6 to 52.0 per 100,000) with seasonality was observed in the state. This was in consonance with a study carried out to study the decadal trends of dengue across all states in the country, wherein Punjab had reported the highest dengue incidence compared to the national median annual incidence of 6.57 per lakh population (Singh et al., 2022). Therefore, the study highlights an urgent need for proactive measures to curb the disease burden in the state.

Additionally, in the study undertaken, exploration of eco-socio-demographic factors and their associations established the strongest relationship of dengue with minimum temperature, cumulative precipitation, relative humidity, and vegetation cover at time lags of around 2–3 months. Such information has policy implications. The health departments can plan prevention and control mechanisms for dengue in advance based on inputs from non-health sector departments monitoring ecological factors such as climatic and environmental data. However, the lag associations for disease forecasting need to be studied in local context as it varies from place to place. A study carried out to estimate the lag effect of climatic variables with dengue at the national level in India found the highest association of temperature and rainfall with dengue occurrence at lags of 3–8 weeks and 9–20 weeks, respectively (Kakarla et al., 2019). In comparison, another study carried out in Brazil found shorter lag associations at 1–2 months for both temperature and rainfall (Lowe et al., 2018). These differences may be attributed to the components of the lag time period, which includes the time taken for the development and reporting of the disease, which is dependent on local ecological and climatic conditions (Farrar and Manson, 2014).

The selection of time stamps is among the most important considerations for time series analysis and should be based on data availability and the amount of time-series long-memory' and noise components (Forecasting: Principles and Practice (3rd Ed), 2021). In the present study, the Hurst coefficient and spectral entropy statistics suggested the selection of monthly intervals for the development of forecasting models. This is another component which needs to be studied in the local context. In previous studies on dengue modelling, time stamps varying from daily (Titus Muurlink et al., 2018), weekly (Kakarla et al., 2019; Phanitchat et al., 2019; Zhang et al., 2019), monthly (Husnina et al., 2019; Jain et al., 2019; Ramadona et al., 2019; Xu et al., 2020), and annual (Stolerman et al., 2019) data have been used. In previous studies to understand RHIS data quality and challenges for its use in research, major determinants for such variations are attributed to the frequency of reporting in health departments, their data capture mechanisms, and data quality characteristics (Deeny and Steventon, 2015; Kumar et al., 2018; Zodpey and Negandhi, 2016).

Exploratory analysis in a parsimonious approach enables researchers to strengthen the plausibility of the forecasting models (Strimbu et al., 2017). The present study explored spatio-temporal associations of dengue with multiple climatic, environmental and socio-demographic factors, and thus proposes a post-preprocessing exploratory framework for research projects using RHIS datasets. Such ESTDA should be undertaken before the development of forecasting models. It is imperative to understand that multiple data analytical approaches have been used for dengue forecasting in the literature, such as time series decomposition analysis(Xu et al., 2020), time series ARMA-based models (Zahirul Islam et al., 2018), generalised spatiotemporal regression (Zheng et al., 2019), and machine learning (Stolerman et al., 2019) among others. However, irrespective of the modelling approach undertaken, the model diagnostics have shown variations and are prone to errors when deliberations on model assumptions and feature selection procedures are inappropriate, highlighting the need for a robust statistical exploratory framework. The present study showed a significant autocorrelation of dengue incidence and a significant seasonal component on STL decomposition. These findings suggest the feasibility of using both approaches for the available RHIS data. Also, ESTDA highlighted the presence of inter-relationships between risk factors and non-linearity in

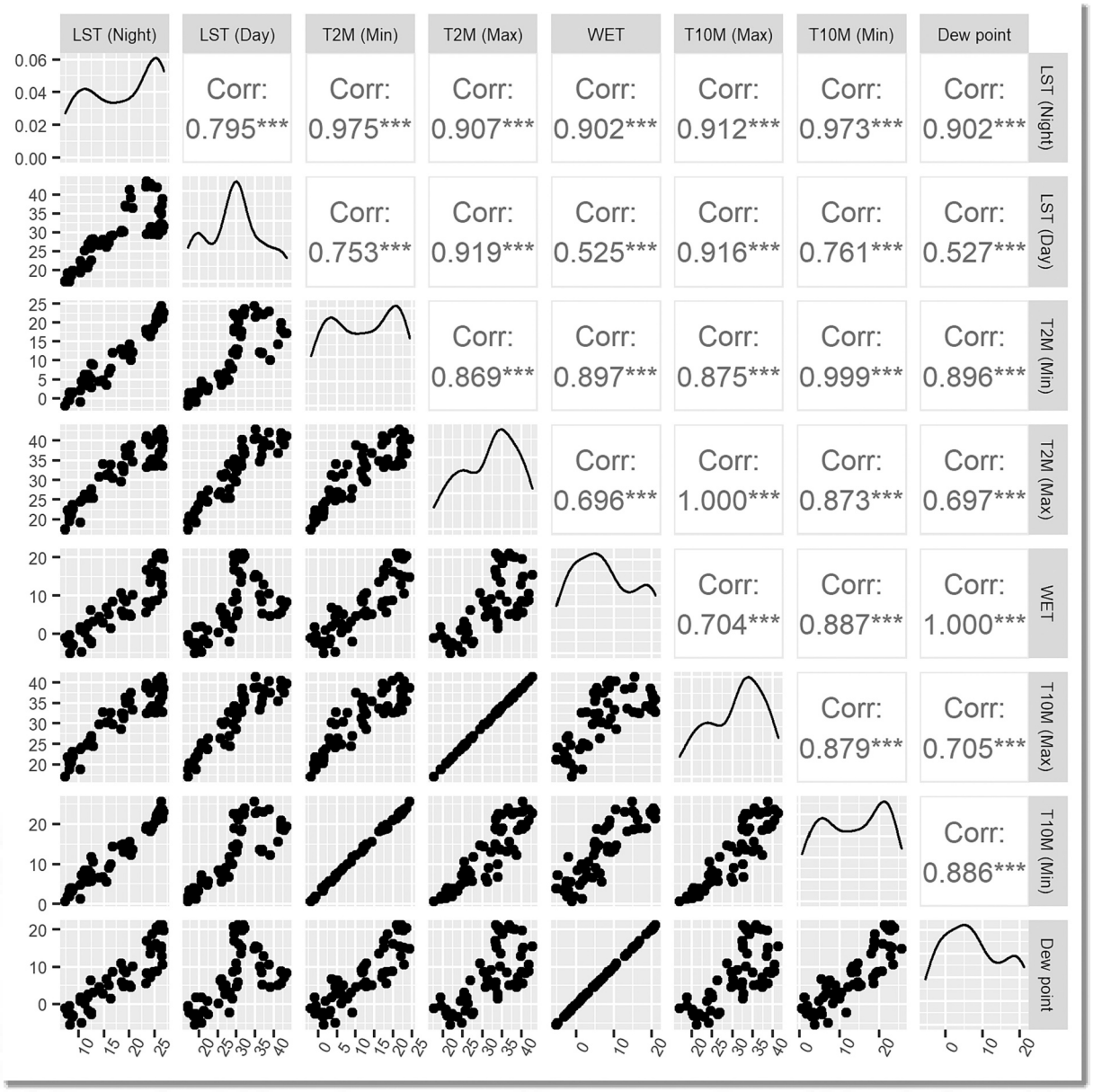


Fig. 3. Correlation matrix between climatic features from multiple data sources.

associations with dengue, violating assumptions for simple linear models. These findings suggest using generalised models and enabling efficient feature selection for future analysis and model development. These findings are similar to previous studies carried out in LMICs in tropical regions (Aswi et al., 2019; de Oliveira-Júnior et al., 2019; Swain et al., 2019; Withanage et al., 2018; Zahirul Islam et al., 2018; Zheng et al., 2019).

In the present study, dengue transmission increased in the northern hilly sub-districts and moved towards a perineal pattern in the southern sub-districts. This can be attributed to the rising temperatures in elevated northern regions and changing ecology due to increased rainfall in the southern areas of the state (positive trends on Seasonal Mann Kendall test). We could not establish this relationship with statistical significance in the present study. Climatic change is a known phenomenon across the globe and is resulting in changing epidemiology of diseases. A study by 'The Energy and Resources Institute' on climatic change and disease dynamics in India has also emphasised the effects of climatic changes in the country and the need for studies inclusive of the climate change perspective for evidence generation (Dogra et al., 2012).

The lack of statistical evidence in the present study may be attributed to the limited time span for which data was collected. This is similar to a previous study undertaken to understand India's climatic associations with dengue occurrence (Kakarla et al., 2019). Larger retrospective dengue RHIS data could not be included in the present study because of the recent advances in disease diagnostics and data collection methods (National Centre for Disease Control, Directorate General of Health Services, 2022). However, studies with longer time duration in future to further evaluate effect of climatic change on dengue occurrence in the region.

Using RHIS and data linkages with non-health sectors using the data science approach and open-source platform to add value to the RHIS till the sub-district level is a novel project in the state. Using established methods in spatial and time series data analytics for ESTDA and futuristic tidy methods for codes/ algorithms provides methodological and scientific strength to the study. Because of misreporting, underreporting, and missed cases, routine data sources are often debated for their quality. Though a concern for the present study, the assumptions of space-time randomness make the analysis undertaken valid for further

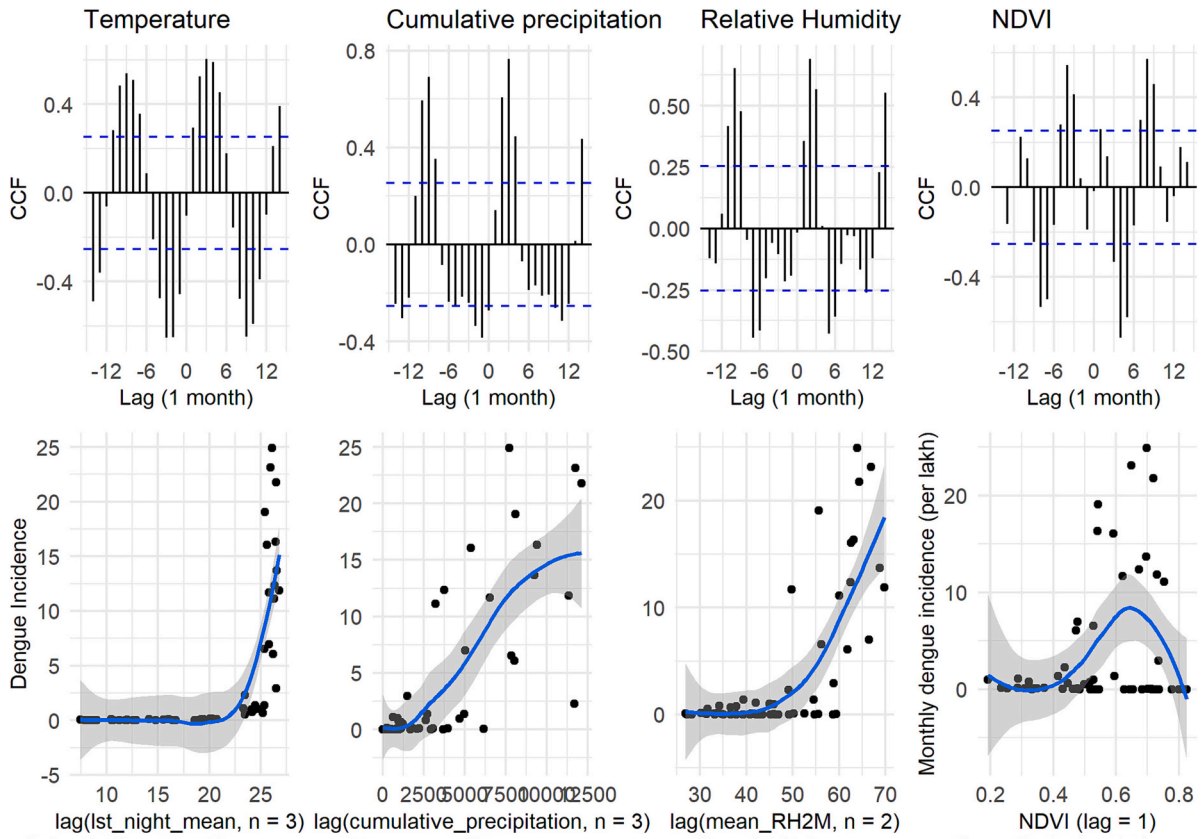


Fig. 4. Time series cross correlation and scatter plots.

research. The patterns found were biologically plausible and in concordance with the literature. Increased reporting of dengue cases may be attributed to multiple factors, including better case detection and reporting rates in the state over a period beyond the scope of the present study. Future work on determining the reasons for the same must provide practical solutions to curb this increasing burden of Dengue in the state.

5. Conclusion

The present study provides evidence and a framework for the exploration of Spatio-temporal associations of dengue with ecological and socio-demographic variables in the local context. The study found a high dengue incidence in the state with seasonal patterns. At the sub-district level, changing epidemiology of dengue was observed, which strengthens the call for climate change policy implementation. A non-linear association of dengue with risk factors was seen at multiple lags and with socio-demographic factors. Also, the study identified significant cross-correlations between risk factors. The study proposes mechanisms by which health departments can use ecological data for the development of dengue forecasting models in LMIC settings. Being open source, reproducible and scalable, the availability and use of these algorithms shall strengthen disease surveillance mechanisms and demonstrates the use of ecological data resources for healthcare systems in resource-constrained settings globally.

Availability of data and algorithms

All the analyses were carried out using open-domain data sources and reproducible codes, which can be shared with readers on reasonable requests. The data from NVBDCP, Punjab, and Sub-district level spatial files were obtained with restricted use, which can be shared only after additional permissions from the state directorate and Punjab Remote Sensing Authority.

Declaration of Competing Interest

None.

Data availability

The authors do not have permission to share data.

References

Aswi, A., Cramb, S.M., Moraga, P., et al., 2019. Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. Epidemiol. Infect. 147, e33 https://doi.org/10.1017/S0950268818002807.
Bivand, R., Pebesma, E.J., Gómez-Rubio, V., 2013. Applied Spatial Data Analysis with R. Second edition. Use R! Springer, New York.
Bruce, P.C., Bruce, A., 2017. Practical Statistics for Data Scientists: 50 Essential Concepts, First edition. O'Reilly, Sebastopol, CA.
Data Science - MeSH - NCBI, 2023. Available at: https://www.ncbi.nlm.nih.gov/mesh/?term=data+science (accessed 7 January 2021).
de Oliveira-Júnior, J.F., Gois, G., da Silva, E.B., et al., 2019. Non-parametric tests and multivariate analysis applied to reported dengue cases in Brazil. Environ. Monit. Assess. 191 (7), 473. https://doi.org/10.1007/s10661-019-7583-0, 7.
De Smith, M.J., Goodchild, M.F., Longley, P.A., 2018. Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools, Sixth edition. Drumlin Security, London.
Deeny, S.R., Steventon, A., 2015. Making sense of the shadows: priorities for creating a learning healthcare system based on routinely collected data. BMJ Qual. Saf. 24 (8), 505–515. https://doi.org/10.1136/bmjqs-2015-004278. BMJ Publishing Group Ltd.
Dengue and severe dengue, 2023. Available at: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue (accessed 1 September 2022).
Directorate General of Health Services, India, 2015. Joint Monitoring Mission Report. Ministry of Health and family Welfare, Govt of India.
Dogra, N., Srivastava, S., Energy and Resources Institute (Eds.), 2012. Climate Change and Disease Dynamics in India. The Energy and Resources Institute, New Delhi.
Farrar, J., Manson, P. (Eds.), 2014. Manson's Tropical Diseases, 23. ed. Elsevier Saunders, Edinburgh. Expertconsult.com.
Forecasting: Principles and Practice (3rd Ed), 2021. Available at: https://otexts.com/fpp3/index.html (accessed 30 July 2022).

Hung, Y.W., Hoxha, K., Irwin, B.R., et al., 2020. Using routine health information data for research in low- and middle-income countries: a systematic review. BMC Health Serv. Res. 20 (1), 790. https://doi.org/10.1186/s12913-020-05660-1.

Husnina, Z., Clements, A.C.A., Wangdi, K., 2019. Forest cover and climate as potential drivers for dengue fever in Sumatra and Kalimantan 2006–2016: a spatiotemporal analysis. Tropical Med. Int. Health. https://doi.org/10.1111/tmi.13248 tmi.13248.

Jain, R., Sontisirikit, S., Iamsirithaworn, S., et al., 2019. Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. BMC Infect. Dis. 19 (1), 272. https://doi.org/10.1186/s12879-019-3874-x, 1.

Jamison, D.T., World Bank and Disease Control Priorities Project, 2006. Disease Control Priorities in Developing Countries, 2nd ed. Oxford University Press; World Bank, New York: Washington, DC.

Kakarla, S.G., Caminade, C., Mutheneni, S.R., et al., 2019. Lag effect of climatic variables on dengue burden in India. Epidemiol. Infect. 147, e170 https://doi.org/10.1017/S0950268819000608.

Kumar, M., Gotz, D., Nutley, T., et al., 2018. Research gaps in routine health information system design barriers to data quality and use in low- and middle-income countries: a literature review. Int. J. Health Plann. Manag. 33 (1), e1–e9. https://doi.org/10.1002/hpm.2447.

Lowe, R., Gasparrini, A., Van Meerbeeck, C.J., et al., 2018. Nonlinear and delayed impacts of climate on dengue risk in Barbados: a modelling study. PLoS Med. 15 (7), e1002613 https://doi.org/10.1371/journal.pmed.1002613.

MEASURE Evaluation, 2022. Routine Health Information Systems. Available at: https://www.measureevaluation.org/our-work/routine-health-information-systems.html (accessed 11 November 2022).

National Centre for Disease Control, Directorate General of Health Services, 2022. Integrated Disease Surveillance Programme(IDSP). Available at: https://idsp.nic.in/index4.php?lang=1&level=0&linkid=313&lid=1592 (accessed 11 November 2022).

National Data Sharing and Accessibility Policy | Department Of Science & Technology, 2022. Available at: https://dst.gov.in/national-data-sharing-and-accessibility-policy-0 (accessed 30 July 2022).

Pebesma, E., 2018. Simple features for R: standardised support for spatial vector data. R J. 10 (1), 439. https://doi.org/10.32614/RJ-2018-009.

Phanitchat, T., Zhao, B., Haque, U., et al., 2019. Spatial and temporal patterns of dengue incidence in northeastern Thailand 2006–2016. BMC Infect. Dis. 19 (1), 743. https://doi.org/10.1186/s12879-019-4379-3, 1.

Ramadona, A.L., Tozan, Y., Lazuardi, L., et al., 2019. A combination of incidence data and mobility proxies from social media predicts the intra-urban spread of dengue in Yogyakarta, Indonesia. In: Werneck, G.L. (Ed.), PLOS Neglected Tropical Diseases, 13. https://doi.org/10.1371/journal.pntd.0007298 (4). 4: e0007298.

Singh, G., Soman, B., Mitra, A., 2021. A Systematic Approach to Cleaning Routine Health Surveillance Datasets: An Illustration Using National Vector Borne Disease Control Programme Data of Punjab, India. arXiv:2108.09963 [cs]. Available at: http://arxiv.org/abs/2108.09963.

Singh, G., Mitra, A., Soman, B., 2022. Development and use of a reproducible framework for spatiotemporal climatic risk assessment and its association with decadal trend of dengue in India. Indian J. Community Med. 47 (1), 50. https://doi.org/10.4103/ijcm.ijcm_862_21.

Stolerman, L.M., Maia, P.D., Kutz, J.N., 2019. Forecasting dengue fever in Brazil: an assessment of climate conditions. PLoS One 14 (8). https://doi.org/10.1371/journal.pone.0220106. Samy AM (ed.). 8: e0220106.

Strimbu, B.M., Amarioarei, A., Paun, M., 2017. A parsimonious approach for modeling uncertainty within complex nonlinear relationships. Ecosphere 8 (9). https://doi.org/10.1002/ecs2.1945.

Swain, S., Bhatt, M., Pati, S., et al., 2019. Distribution of and associated factors for dengue burden in the state of Odisha, India during 2010–2016. Infect. Dis. Poverty 8 (1), 31. https://doi.org/10.1186/s40249-019-0541-9, 1.

Titus Muurlink, O., Stephenson, P., Islam, M.Z., et al., 2018. Long-term predictors of dengue outbreaks in Bangladesh: a data mining approach. Infect. Dis. Model. 3, 322–330. https://doi.org/10.1016/j.idm.2018.11.004.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley Series in Behavioral Science. Addison-Wesley Pub. Co, Reading, Mass.

van der Aalst, W.M.P. (Ed.), 2016. Process Mining: Data Science in Action, 2nd ed. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-49851-4. Imprint: Springer.

Wickham, H., Averick, M., Bryan, J., et al., 2019. Welcome to the Tidyverse. J. Open Source Software 4 (43), 1686. https://doi.org/10.21105/joss.01686.

Withanage, G.P., Viswakula, S.D., Nilmini Silva Gunawardena, Y.I., et al., 2018. A forecasting model for dengue incidence in the District of Gampaha, Sri Lanka. Parasit. Vectors 11 (1), 262. https://doi.org/10.1186/s13071-018-2828-2.

Xiao, Z., Xu, X., Xing, H., et al., 2021. RTFN: a robust temporal feature network for time series classification. Inf. Sci. 571, 65–86. https://doi.org/10.1016/j.ins.2021.04.053.

Xing, H., Xiao, Z., Qu, R., et al., 2022a. An efficient federated distillation learning system for multitask time series classification. IEEE Trans. Instrum. Meas. 71, 1–12. https://doi.org/10.1109/TIM.2022.3201203.

Xing, H., Xiao, Z., Zhan, D., et al., 2022b. SelfMatch: robust semisupervised time-series classification with self-distillation. Int. J. Intell. Syst. 37 (11), 8583–8610. https://doi.org/10.1002/int.22957.

Xu, Z., Bambrick, H., Yakob, L., et al., 2020. High relative humidity might trigger the occurrence of the second seasonal peak of dengue in the Philippines. Sci. Total Environ. 708, 134849 https://doi.org/10.1016/j.scitotenv.2019.134849.

Zahirul Islam, M., Rutherford, S., Phung, D., et al., 2018. Correlates of climate variability and dengue fever in two metropolitan cities in Bangladesh. Cureus. https://doi.org/10.7759/cureus.3398.

Zhang, Qin, Chen, Y., Fu, Y., et al., 2019. Epidemiology of dengue and the effect of seasonal climate variation on its dynamics: a spatio-temporal descriptive analysis in the Chao-Shan area on China's southeastern coast. BMJ Open 9 (5). https://doi.org/10.1136/bmjopen-2018-024197, 5: e024197.

Zheng, L., Ren, H.-Y., Shi, R.-H., et al., 2019. Spatiotemporal characteristics and primary influencing factors of typical dengue fever epidemics in China. Infect. Dis. Poverty 8 (1), 24. https://doi.org/10.1186/s40249-019-0533-9, 1.

Zodpey, S., Negandhi, H., 2016. Improving the quality and use of routine health data for decision-making. Indian J. Public Health 60 (1), 1. https://doi.org/10.4103/0019-557X.177248.

APPENDIX E – REPRODUCIBLE ALGORITHMS

ALGORITHM FOR MODIS DATA EXTRACTION.

Enter details

```
# Enter polygon location
path_polygon <- "file_name_and_path.rds"
# Enter HDF file location: Ensure no files other than HDF MOD11C2
files are present in the directory
path_mod11c2 <- "file_path"
# Enter output file name
output_file_name <- "file_output_name_and_path.rds"
```

```
library(tidyverse)
library(sf)
library(raster)
library(gdalUtils)
```

Load polygon file

```
polygon_punjab <- read_sf(path_polygon)
```

Load files

```
files <- fs::dir_ls(path_mod11c2)
```

Loop for data extraction.

```
file = NULL
hdf_layer = NULL
extracted_data = NULL
for(i in seq_along(files)){
  file = files[[i]]
  file_parts <- unlist(strsplit(basename(file), "\\."))
  date_string <- file_parts[[2]]
  year <- as.numeric(substr(date_string, 2, 5))
  day <- as.numeric(substr(date_string, 6, 8))
  sds <- MODIS::getSds(file)
  hdf_layer <- raster::raster(rgdal::readGDAL(sds$SDS4gdal[6], as.is
= TRUE))
  #### Crop raster layer to polygon extent ####
  polygon_punjab <- st_transform(polygon_punjab,
                                 st_crs(hdf_layer))
  ## Apply scale factor and offset.
  scale_factor <- 0.02
  hdf_layer <- (hdf_layer * scale_factor)-272
  ## Cropping raster file to punjab bbbox
  lst_punjab_date <- crop(hdf_layer,
                          polygon_punjab,
                          snap = "out")
  # Extract polygon wise data
  extracted_data[[i]] <-
exactextractr::exact_extract(lst_punjab_date,
```

```
                                                    polygon_punjab,
                                                    c("min",
                                                      "max",
                                                      "mean",
                                                      "median",
                                                      "stdev",
                                                      "mode"),
                                                    append_cols =
"block_district")
  names(extracted_data[[i]]) <- str_c("lst_night",
                                      names(extracted_data[[i]]),
                                      year,
                                      day,
                                      sep = "_")

}
```

Combine all extracted data files

```
combined_lst <- do.call(cbind, extracted_data)
```

Save file

```
write_rds(combined_lst,
          output_file_name)
```

ALGORITHM FOR IMERG DATA EXTRACTION.

Enter details

```
path_files <- "file_path"
# Polygon file location
path_polygon <- "file_name_location.shp"
# Output file name
output_file_name <- "output_file_name_and_path.rds"
```

List of files

```
files = list.files(path_files, pattern = ".nc4$")
```

Create folder for raster files

```
suppressWarnings(dir.create(paste(path_files,"imerg_raster",sep="/")))
```

Data extraction loop

```
library(tidyverse)
library(sf)
library(ncdf4)
library(raster)
#Open ncdf file
file = NULL
i = NULL
extracted_data = NULL
for (i in seq_along(files)){
file <- files[i]
nc<-ncdf4::nc_open(stringr::str_c(path_files, file))
#getting the y values (longitudes in degrees east)
nc.long.IMERG<-ncdf4::ncvar_get(nc,nc$dim[[1]])
#getting the x values (latitudes in degrees north)
nc.lat.IMERG<-ncdf4::ncvar_get(nc,nc$dim[[2]])
#extract data
data<-ncdf4::ncvar_get(nc,'precipitationCal')
#reorder the rows
data<-data[ nrow(data):1, ]
ncdf4::nc_close(nc)
#convert to raster
imerg_date <-
raster::raster(x=as.matrix(data),xmn=nc.long.IMERG[1],xmx=nc.long.IMERG
[NROW(nc.long.IMERG)],ymn=nc.lat.IMERG[1],ymx=nc.lat.IMERG[NROW
(nc.lat.IMERG)],crs=sp::CRS('+proj=longlat +datum=WGS84'))
rm(nc.long.IMERG,nc.lat.IMERG,nc,data)
# Save the extracted raster files
date <- stringr::str_sub(file,22,29)
```

```
filename_raster <- stringr::str_c(path_files,
"imerg_raster","/",date,"imerg_raster.tif")
raster::writeRaster(
  imerg_date,
  filename_raster)
rm(filename_raster, file)
# Crop raster layer to polygon extent
## Load polygon file
polygon_punjab <- read_sf(path_polygon)
names(polygon_punjab) <- epitrix::clean_labels(names(polygon_punjab))
## Transform crs of polygon as raster
polygon_punjab <- st_transform(polygon_punjab,
                               st_crs(imerg_date))

polygon_punjab$block_district <-
stringr::str_c(polygon_punjab$new_block, polygon_punjab$new_dist, sep =
", ")
## Cropping raster file to punjab bbbox
imerg_punjab_date <- crop(imerg_date,
                          polygon_punjab,
                          snap = "out")
# Extract polygon wise data
extracted_data[[i]] <- exactextractr::exact_extract(imerg_punjab_date,
                polygon_punjab,
                "sum",
                append_cols = "block_district")
names(extracted_data[[i]]) <- str_c(names(extracted_data[[i]]),
"imerg", date, sep = "_")
}
```

Combine extracted data

```
combined_imerg <- do.call(cbind, extracted_data)
```

Save extracted data

```
write_rds(combined_imerg,
          output_file_name)
```

ALGORITHM FOR NASAPOWER API DATA EXTRACTION.

Determine location details of the study area.

```
library(nasapower)
library(tidyverse)
library(sf)
#Load data
shp <- read_sf("file_location.shp")
```

Convert polygon CRS to match CRS of NASA POWER data.

```
shp <- st_transform(shp, 4326)
```

Determine bounding box details for the study area.

```
st_bbox(shp)
```

Get POWER data

```
daily_region_ag <- get_power(
  community ="AG",
  lonlat =c(73,29,77,33), # specify bounding box details
  pars =c("PRECTOT", # precipitation
          "QV2M", # Specific Humidity at 2 Meters
          "RH2M", # Relative Humidity at 2 Meters
          "T10M_MAX", # Maximum Temperature at 10 Meters
          "T10M_MIN", # Minimum Temperature at 10 Meters
          "T2M", # Temperature at 2 Meters
          "T2MDEW", # Dew/Frost Point at 2 Meters
          "T2MWET", # Wet Bulb Temperature at 2 Meters
          "T2M_MAX", # Maximum Temperature at 2 Meters
          "T2M_MIN", # Minimum Temperature at 2 Meters
          "WS10M", # Wind Speed at 10 Meters
          "WS2M" # Wind Speed at 2 Meters
          ),
  dates = c("2014-01-01","2020-01-01"), # Specify dates
  temporal_average ="DAILY")
```

Save file

```
write_rds(daily_region_ag,
          "file_name_and_path.rds")
```

ALGORITHM FOR SPATIAL AUTOCORRELATION ANALYSIS

```
library(tidyverse)
library(lubridate)
library(flextable)
library(gtsummary)
library(sf)
library(spdep)
```

Annual Moran's I

```
df_block <- readRDS(here::here("file_name_and_path.rds"))
# Create block level monthly time series dataset
df <- df_block %>%
  group_by(year(expected_date_test_v2),
           # month(expected_date_test_v2),
           NEW_BLOCK, .add = T) %>%
  summarise(cases = sum(cases),
            population = mean(population),
            cumulative_rainfall = sum(PRECTOT),
            mean_temp = mean(T2MWET),
            mean_relative_humidity = mean(RH2M),
            .groups = "keep") %>%
  ungroup()
shp_blocks <- readRDS(here::here("file_name_and_path.rds"))
```

Crude dengue incidence rates (per lakh population)

```
df <- df %>%
  dplyr::mutate(crude_dengue_incidence =
           round(cases/population*100000, 2))
```

Moran's I

```
# Creating a neighbor list from polygon list
neighbor_list_q <-
  shp_blocks %>%
  poly2nb(queen = T)
# Spatial weights
neighbor_weights_q_w <-
  neighbor_list_q %>%
  nb2listw(style = "W")
# Moran's I
df_moran <- left_join(shp_blocks,
                      df)

output_moran <- df_moran %>%
  split(.$`year(expected_date_test_v2)`)

poss_moran = possibly(.f = ~moran.test(.$crude_dengue_incidence,
                  listw = neighbor_weights_q_w,
```

```
                    na.action=na.omit),
                    otherwise = "Error")

moran_global <- output_moran %>%
  map(.,poss_moran)

# statistic values
moran_value <- possibly(.f = ~.$estimate[[1]],
                    otherwise = "error")
moran_stat <- moran_global %>%
  map_df(., moran_value) %>%
  map_df(., as.character) %>%
  pivot_longer(cols = 1:dim(.)[[2]],
               values_to = "stat")

# Expected values
moran_expected <- possibly(.f = ~.$estimate[[2]],
                    otherwise = "error")
moran_expected <- moran_global %>%
  map_df(., moran_expected) %>%
  map_df(., as.character) %>%
  pivot_longer(cols = 1:dim(.)[[2]],
               values_to = "expected")

# z scores
moran_zscore <- possibly(.f = ~.$statistic,
                    otherwise = "error")
moran_zscore <- moran_global %>%
  map_df(., moran_zscore) %>%
  map_df(., as.character) %>%
  pivot_longer(cols = 1:dim(.)[[2]],
               values_to = "z score")
moran_zscore <- moran_zscore %>%
  mutate(name = moran_expected$name)
# p values
p_value <- possibly(.f = ~.$p.value,
                    otherwise = "error")
significance <- moran_global %>%
  map_df(., p_value) %>%
  map_df(., as.character) %>%
  pivot_longer(cols = 1:dim(.)[[2]],
               values_to = "p_value") %>%
  dplyr::mutate(Significance =
           case_when(
             as.numeric(p_value) <= 0.05 ~ "Significant",
             T ~ "Not Significant"
           ))

# Variance
```

```
moran_variance <- possibly(.f = ~.$estimate[[3]],
                    otherwise = "error")
moran_variance <- moran_global %>%
  map_df(., moran_variance) %>%
  map_df(., as.character) %>%
  pivot_longer(cols = 1:dim(.)[[2]],
               values_to = "variance")

# Join data
significance<- left_join(moran_stat,
                         significance)
significance <- left_join(moran_expected, significance)
significance <- left_join(moran_zscore, significance)
significance <- left_join(moran_variance, significance)
# save table
# write.csv(significance, "annual_moran.csv")
```

Monthly Moran's I

Load and pre-process data

```
rm(list = ls())
df_block <-
readRDS(here::here("data","final_block_analysis_datasets.rds"))
# Create block level monthly time series dataset
df <- df_block %>%
  group_by(year(expected_date_test_v2),
           month(expected_date_test_v2),
           NEW_BLOCK, .add = T) %>%
  summarise(cases = sum(cases),
            population = mean(population),
            cumulative_rainfall = sum(PRECTOT),
            mean_temp = mean(T2MWET),
            mean_relative_humidity = mean(RH2M),
            .groups = "keep") %>%
  ungroup()
rm(df_block)
shp_blocks <- readRDS(here::here("data", "final_shp_blocks.rds"))
```

Crude dengue incidence rates (per lakh population)

```
df <- df %>%
  dplyr::mutate(crude_dengue_incidence =
           round(cases/population*100000, 2))
```

Moran's I

```
library(spdep)
# Creating a neighbor list from polygon list
neighbor_list_q <-
  shp_blocks %>%
  poly2nb(queen = T)
```

```r
# neighbor_weights_q_b
neighbor_weights_q_w <-
  neighbor_list_q %>%
  nb2listw(style = "W")
neighbor_weights_q_w
# Moran's I
df_moran <- left_join(shp_blocks,
                      df)
df_moran$year_month <- str_c(df_moran$`year(expected_date_test_v2)`,

df_moran$`month(expected_date_test_v2)`,
                             sep = "_")

output_moran <- df_moran %>%
  split(.$year_month)

poss_moran = possibly(.f = ~moran.test(.$crude_dengue_incidence,
                  listw = neighbor_weights_q_w,
                  na.action=na.omit),
                  otherwise = "Error")

moran_global <- output_moran %>%
  map(.,poss_moran)
moran_global$`2015_10`$estimate[[1]]
# statistic values
moran_value <- possibly(.f = ~.$estimate[[1]],
                    otherwise = "error")
moran_stat <- moran_global %>%
  map_df(., moran_value) %>%
  map_df(., as.character) %>%
  pivot_longer(cols = 1:dim(.)[[2]],
               values_to = "stat")

# Expected values
moran_expected <- possibly(.f = ~.$estimate[[2]],
                    otherwise = "error")
moran_expected <- moran_global %>%
  map_df(., moran_expected) %>%
  map_df(., as.character) %>%
  pivot_longer(cols = 1:dim(.)[[2]],
               values_to = "expected")

# z scores
moran_zscore <- possibly(.f = ~.$statistic,
                    otherwise = "error")
moran_zscore <- moran_global %>%
  map_df(., moran_zscore) %>%
  map_df(., as.character) %>%
```

```
  pivot_longer(cols = 1:dim(.)[[2]],
               values_to = "z score")
moran_zscore <- moran_zscore %>%
  mutate(name = moran_expected$name)
# p values
p_value <- possibly(.f = ~.$p.value,
                    otherwise = "error")
significance <- moran_global %>%
  map_df(., p_value) %>%
  map_df(., as.character) %>%
  pivot_longer(cols = 1:dim(.)[[2]],
               values_to = "p_value") %>%
  dplyr::mutate(Significance =
           case_when(
             as.numeric(p_value) <= 0.05 ~ "Significant",
             T ~ "Not Significant"
           ))
# Variance
moran_variance <- possibly(.f = ~.$estimate[[3]],
                    otherwise = "error")
moran_variance <- moran_global %>%
  map_df(., moran_variance) %>%
  map_df(., as.character) %>%
  pivot_longer(cols = 1:dim(.)[[2]],
               values_to = "variance")
```

ALGORITHM FOR EMERGING HOTSPOT ANALYSIS

Load libraries and data

```
library(tidyverse)
library(sf)
library(rgeoda)
library(spdep)
library(lubridate)
library(Kendall)
shp_blocks <- readRDS(here::here("file_name_and_path.rds"))
df_block <- readRDS(here::here("file_name_and_path.rds"))
```

Create monthly ts

```
df <- df_block %>%
  group_by(year = year(expected_date_test_v2),
           month = month(expected_date_test_v2),
           NEW_BLOCK, .add = T) %>%
  summarise(cases = sum(cases),
            population = mean(population),
            .groups = "keep") %>%
  ungroup() %>%
  dplyr::mutate(crude_dengue_incidence =
           round(cases/population*100000, 2))
```

Create a neighbor list from polygon list

```
neighbor_list_q <-
  shp_blocks %>%
  poly2nb(queen = T)
```

Create Spatial weights with self included for G star

```
neighbor_weights_q_w <-
   nb2listw(include.self(neighbor_list_q),
 style="W")
df_G <- left_join(shp_blocks,
                        df)
df_G$year_month <- str_c(df_G$year,
                           df_G$month,
                           sep = "_")
```

Calculate Gi* values

```
output_G <- df_G %>%
  split(.$year_month)

local_G_fn <- function(df){
  localG(df$crude_dengue_incidence,
       listw = neighbor_weights_q_w)
}
```

```
G_star <- map_df(output_G, local_G_fn)

df <- cbind(new_block = output_G$`2015_1`$new_block,
            G_star)
```

Tidy data

```
df <- df %>%
  pivot_longer(cols = 2:61,
               names_to = "year_month",
               values_to = "g_z_score")
```

Trend analysis

```
selected_block = as.list(unique(df$new_block))

kendall_map_fn <- function(selected_block){
  df <- df %>%
  filter(new_block == selected_block) %>%
    mutate(timeline = ym(year_month)) %>%
    arrange(timeline)
  x = SeasonalMannKendall(ts(df$g_z_score))
  kendall_df <- tibble(new_block = selected_block,
       S =
         round(as.numeric(as.character(x)[[3]]),3),
      kendall_p =
         round(as.numeric(as.character(x)[[2]]),3))
}

kendall_df <- map_df(selected_block, kendall_map_fn)
```

Add Hotspot and coldspot category

```
df <- df %>%
  mutate(
    cat_g_star = case_when(
      g_z_score >= 1.96 ~ "Hotspot",
      g_z_score <= 1.96 ~ "Coldspot",
      TRUE ~ "Not Significant"))
```

Emerging Hotspot categorization

```
df_hotspots_trend <- df_hotspots_trend %>%
  mutate(
    emerging_hotspot = case_when(
      hot_2015+hot_2016+hot_2017+hot_2018+hot_2019 >=4 &
        trend_cat == "Positive trend" ~ "Persistent and
Intensifying",
      hot_2015+hot_2016+hot_2017+hot_2018+hot_2019 >=4 &
        trend_cat == "Not significant" ~ "Persistent",
      hot_2015+hot_2016+hot_2017+hot_2018+hot_2019 >=4 &
        trend_cat == "Negative trend" ~ "Persistent and
Diminishing",
```

```
      hot_2015+hot_2016 == 0 &
       hot_2017+hot_2018+hot_2019 >=2 ~ "Emerging",
      hot_2019 == 1 &
       hot_2015+hot_2016+hot_2017+ hot_2018 == 0 ~ "New",
      hot_2015+hot_2016+hot_2017+hot_2018+hot_2019 ==3 ~
"Oscillating",
      hot_2015+hot_2016+hot_2017 >=2 &
        hot_2018+hot_2019 ==0 ~ "Historical",
      hot_2015+hot_2016+hot_2017+hot_2018+hot_2019 >=1 ~ "Sporadic",
      TRUE ~ "Not categorised"))
```

ALGORITHM FOR GLMs

Load library and data

```
library(tidyverse)
library(lubridate)
library(MASS)
library(pscl)
library(gtsummary)
library(here)
df <- read_rds(here("file_name_and_path.rds"))
```

Quasi poisson

```
model_clim = glm(cases ~ NEW_BLOCK +
                         lag3_temp +
                         lag3_ppt +
                         lag2_rel_humidity+
                         lag1_wind_speed +
                         lag1_ndvi +
                         urbanization +
                         elevation +
                         household_density +
                         female_literacy_rate +
                         offset(log(population)),
                       data = df,
                     family = "quasipoisson")
```

Negative binomial model

```
model_clim = glm.nb(cases ~ NEW_BLOCK +
                         lag3_temp +
                         lag3_ppt +
                         lag2_rel_humidity+
                         lag1_wind_speed +
                         lag1_ndvi +
                         urbanization +
                         elevation +
                         household_density +
                         female_literacy_rate +
                         offset(log(population)),
                       data = df)
```

ALGORITHM FOR GAMMs BASED FORECASTING

Load libraries

```
library(tidyverse)
library(lubridate)
library(mgcv)
library(here)
```

Load and pre-process data

```
df <- readRDS(here("file_name_and_path.rds"))
```

T2P2H2 model

```
model_clim = gam(cases ~ NEW_BLOCK+
                   s(lag2_temp,    bs = "cr") +
                   s(lag2_ppt,   bs = "cr") +
                   s(lag2_rel_humidity,   bs = "cr")+
                   s(lag1_wind_speed,   bs = "cr") +
                   s(lag1_ndvi,   bs = "cr")+
                   s(urbanization,   bs = "cr")+
                   s(slope,    bs = "cr")+
                   s(household_density, bs = "cr") +
                   s(female_literacy_rate,   bs = "cr")+
                   s(area,   bs = "cr")+
                   offset(log(population)),
                 data = df,
                 method = "REML",
                 family = nb)
```

Final model with categorical data

```
model_clim = gam(cases ~ s(NEW_BLOCK, bs = "re")+
                   s(lag2_temp,   k = 50, bs = "cr") +
                   s(lag2_ppt, k = 50, bs = "cr") +
                   s(lag2_rel_humidity,k = 50, bs = "cr")+
                   s(lag1_wind_speed,   k = 50, bs = "cr") +
                   s(lag1_ndvi, k = 50, bs = "cr")+
                   urban_cat +
                   slope_cat +
                   literacy_cat +
                   s(area,   bs = "cr")+
                   s(month, bs = "cc")+
                   offset(log(population)),
                 data = train,
                 method = "REML",
                 family = nb,
                 knots = list(month = c(1, 12)))
```

Forecasting

```
train <- df %>%
  filter(month_id <=48)
test <- df %>%
  filter(month_id > 48)
# Predict
predicted <- predict.gam(model_clim,
            newdata = test,
            type = "response",
            se.fit = T)

test <- cbind(test, predicted$fit)
test <- rename(test,
       "pred" = "predicted$fit")

se = predicted$se.fit

test <- cbind(test, se)
test <- test %>%
  mutate(upperCI = (pred + (2*se))) %>%
  mutate(lowerCI = (pred - (2*se)))
rmse <- sqrt(mean((test$cases - test$pred)^2))
rmse
#RMSE
check_state <- test %>%
  mutate(state = "Punjab") %>%
  group_by(state, month, .add = T) %>%
  summarise(cases = sum(cases),
            pred = sum(pred), .groups = "keep")
rmse_state <- sqrt(mean((check_state$cases - check_state$pred)^2))
rmse_state

## District wise
check_df <- test %>%
  group_by(month_id, district, .add = T) %>%
  summarise(cases = sum(cases),
            pred = sum(pred),
            upperCI = sum(upperCI),
            lowerCI = sum(lowerCI),
            timeline = max(timeline),
            .groups = "keep")

# Accuracy calculations
check_df %>% ungroup() %>%
  dplyr::mutate(
    accuracy = factor(ifelse(
      cases >= lowerCI &
       cases <= upperCI , "Within 95",
      "Not accurate"
```

```
    )
  )) %>%
  select(accuracy, timeline) %>%
  gtsummary::tbl_summary(by = timeline)

## Block wise
check_df <- test %>%
  group_by(month_id, NEW_BLOCK, .add = T) %>%
  summarise(cases = sum(cases),
            pred = sum(pred),
            upperCI = sum(upperCI),
            lowerCI = sum(lowerCI),
            timeline = max(timeline),
            .groups = "keep")
# Accuracy calculations
check_df %>% ungroup() %>%
  dplyr::mutate(
    accuracy = factor(ifelse(
      cases >= lowerCI &
       cases <= upperCI , "Within 95",
      "Not accurate"
    )
  )) %>%
  select(accuracy, timeline) %>%
  gtsummary::tbl_summary(by = timeline)
```

ALGORITHM FOR HEIRARCHIAL FORECASTING

Create a block level monthly ts object

```
library(tidyverse)
library(fpp3)
df_block <- readRDS(here::here("file_name_and_path.rds"))

ts_month <- df_block %>%
   mutate(month = yearmonth(expected_date_test_v2)) %>%
  dplyr::select(-expected_date_test_v2) %>%
  group_by(month, NEW_BLOCK, NEW_DIST, .add = T) %>%
  summarise(NEW_BLOCK = unique(NEW_BLOCK),
            NEW_DIST = unique(NEW_DIST),
            cases = sum(cases),
            month = unique(month),
            .groups = "keep") %>%
  as_tsibble(key = c(NEW_BLOCK, NEW_DIST),
             index = month) %>%
  relocate(month) %>% ungroup()
```

Add heirarchial levels

```
ts <- ts_month %>%
  aggregate_key(NEW_DIST/NEW_BLOCK, Count = sum(cases))
```

Obtain Forecast values using ARIMA

```
month_selected = NULL
forecast_fn <- function(month_selected){
  # Create model
  heirarchial_model <- ts %>%
  dplyr::filter(ym(month) <
                  lubridate::dmy(month_selected)) %>%
  model(base = ARIMA(Count))
  # Fit models
  fit <- heirarchial_model %>%
  reconcile(
    bu = bottom_up(base),
    td = top_down(base),
    mo = middle_out(base),
    ols = min_trace(base, method = "ols"),
    mint = min_trace(base, method = "mint_shrink"),
  )
  # Forecast
  fc <- fit %>% forecast(h = 2)
  fc <- fc %>% filter(month(month) != month(dmy(month_selected)))
}
```

APPENDIX F – SUPPLEMENTARY TABLES

Local Moran's Sensitivity Analysis

LISA	List of Blocks		
	p = 0.01	p = 0.05	p = 0.1
August 2015			
High-High	Aur, Bathinda, Jhunir, Maur, Nathana, Talwandi Sabo, Goiniana	Bathinda, Jhunir, Maur, Nathana, Talwandi Sabo, Goiniana	Bathinda, Nathana
Low-Low	Batala, Khadur Sahib, Kot Ise Khan, Nadala, Dirba	Batala	NULL
Low-High	Balachaur, Banga, Garh Shankar, Kot Bhai At Gidderbaha, Jaitu, Ludhiana Ii, Machhiwara, Sangat, Saroya	Banga, Kot Bhai At Gidderbaha, Jaitu, Sangat, Saroya	Kot Bhai At Gidderbaha, Jaitu
High-Low	Dhar Kalan, Kapurthala, Khanna, Patiala	Dhar Kalan, Kapurthala, Khanna, Patiala	Dhar Kalan, Kapurthala, Khanna, Patiala
September 2015			
High-High	Amloh, Bathinda, Bhawanigarh, Nathana, Sirhind, Talwandi Sabo, Goiniana	Nathana, Talwandi Sabo, Goiniana	Nathana, Talwandi Sabo, Goiniana
Low-Low	Ajnala, Andana, Bhunarheri, Chohla Sahib, Dasuya, Fazilka, Gurdaspur, Lohian, Nur Mahal, Patran, Shahkot	Ajnala, Bhunarheri	Ajnala
Low-High	Adampur, Maloud, Doraha, Ghanaur, Khamanon, Kot Bhai At Gidderbaha, Malerkotla, Maur, Nabha, Samana, Samrala, Sanaur, Sangat, Saroya, Shambu Kalan	Kot Bhai At Gidderbaha, Maur, Nabha, Sangat, Shambu Kalan	Maur, Sangat
High-Low	Batala, Tarn Taran, Verka	Batala, Tarn Taran, Verka	Batala, Tarn Taran
October 2015			
High-High	Amloh, Bassi Pathanan, Bhikhi, Khera, Maur, Rampura, Sangat, Sirhind, Talwandi Sabo, Goiniana	Amloh, Bassi Pathanan, Bhikhi, Khera, Maur, Sangat, Sirhind, Talwandi Sabo	Maur
Low-Low	Ajnala, Andana, Batala, Bhunarheri, Chohla Sahib, Dasuya, Dina Nagar, Fatehgarh Churian, Fazilka, Gurdaspur, Jalalabad, Khuian Sarwar, Lohian, Mukerian, Qadian, Rayya, Shahkot,	Ajnala, Andana, Batala, Bhunarheri, Dasuya, Gurdaspur, Rayya, Sujanpur, Mehtapur	Ajnala, Andana, Batala, Bhunarheri

	Sujanpur, Sultanpur Lodhi, Mehtapur		
Low-High	Adampur, Bhawanigarh, Ghanaur, Jhunir, Khamanon, Kot Bhai At Gidderbaha, Malerkotla, Nabha, Nathana, Samana, Sanaur, Shambu Kalan	Bhawanigarh, Ghanaur, Nabha, Sanaur, Shambu Kalan	Bhawanigarh, Nabha, Shambu Kalan
High-Low	Tarn Taran, Verka	Tarn Taran, Verka	Tarn Taran
November 2015			
High-High	Amloh, Barnala, Bassi Pathanan, Bhikhi, Jhunir, Maur, Rampura, Sangrur, Sirhind	Amloh, Bhikhi, Jhunir, Maur, Rampura	Bhikhi, Maur, Rampura
Low-Low	Ajnala, Batala, Bhikhiwind, Bhunarheri, Chohla Sahib, Dasuya, Dina Nagar, Gurdaspur, Jalalabad, Khadur Sahib, Khuian Sarwar, Malout, Mukerian, Narot Jaimal Singh, Naushera Pannuan, Gharota, Qadian, Rayya, Shahkot, Sri Hargobindpur, Sujanpur, Sultanpur Lodhi, Tarn Taran, Arniwal, Pathankot	Ajnala, Batala, Bhunarheri, Dasuya, Malout, Narot Jaimal Singh, Rayya, Sujanpur, Tarn Taran, Pathankot	Dasuya, Rayya, Tarn Taran
Low-High	Bhawanigarh, Budhlada, Khamanon, Khera, Malerkotla, Mamdot, Nabha, Sher Pur, Talwandi Sabo, Shambu Kalan	Budhlada, Khera, Nabha, Shambu Kalan	Budhlada, Nabha
High-Low	Rupnagar, Verka	Rupnagar, Verka	NULL
December 2015			
High-High	Amloh, Budhlada, Jhunir, Khera, Maur, Nihal Singh Wala, Sidhwan Bet, Sirhind	Budhlada, Jhunir, Maur	Budhlada, Jhunir, Maur
Low-Low	Batala, Dasuya, Dhariwal, Malout, Nadala, Narot Jaimal Singh, Qadian, Rayya, Sri Hargobindpur, Arniwal, Pathankot	Batala, Rayya	NULL
Low-High	Bagha Purana, Bhawanigarh, Bhikhi, Ghanaur, Nabha, Rajpura, Rampura, Samana, Sanaur, Sher Pur, Shambu Kalan	Bhikhi, Ghanaur, Nabha, Rampura, Shambu Kalan	Bhikhi, Rampura
High-Low	Mukerian, Rupnagar, Sultanpur Lodhi	Mukerian, Sultanpur Lodhi	Mukerian
August 2016			
High-High	Aur, Derabassi, Faridkot, Kharar, Majri, Rajpura, Sanaur	Derabassi, Majri, Rajpura	Majri, Rajpura

Low-Low	Abohar, Bhagta Bhai Ka, Jagraon, Kot Ise Khan, Lambi, Lohian, Mahal Kalan, Malout, Moga I, Nihal Singh Wala, Nur Mahal, Phul, Raikot, Sangat, Sehna, Arniwal, Lehra Gaga	Abohar, Bhagta Bhai Ka, Lambi, Mahal Kalan, Moga I, Nihal Singh Wala, Nur Mahal, Phul, Raikot, Sehna	NULL
Low-High	Bagha Purana, Balachaur, Banga, Bassi Pathanan, Ghanaur, Khera, Jaitu, Morinda, Saroya, Shambu Kalan	Balachaur, Khera, Morinda, Shambu Kalan	Khera, Morinda
High-Low	Shahkot, Sujanpur	Shahkot, Sujanpur	Shahkot
September 2016			
High-High	Barnala, Bassi Pathanan, Derabassi, Dhuri, Kharar, Majri, Rajpura	Barnala, Derabassi, Dhuri, Majri, Rajpura	Barnala, Dhuri
Low-Low	Abohar, Bathinda, Bhagta Bhai Ka, Chohla Sahib, Dasuya, Fazilka, Hajipur, Jalalabad, Khuian Sarwar, Kot Bhai At Gidderbaha, Lambi, Lohian, Malout, Nathana, Nur Mahal, Sangat, Shahkot, Arniwal, Goiniana, Mehtapur, Lehra Gaga	Abohar, Bathinda, Bhagta Bhai Ka, Fazilka, Jalalabad, Khuian Sarwar, Kot Bhai At Gidderbaha, Lambi, Malout, Nur Mahal, Sangat, Shahkot, Arniwal	Abohar, Bathinda, Fazilka, Jalalabad, Kot Bhai At Gidderbaha, Lambi, Malout, Arniwal
Low-High	Bhawanigarh, Chogawan, Harsha Chhina, Khera, Malerkotla, Ahemdagarh, Morinda, Shambu Kalan	Khera, Malerkotla, Morinda	NULL
High-Low	Moga I	Moga I	Moga I
October 2016			
High-High	Balachaur, Bassi Pathanan, Chamkaur Sahib, Derabassi, Kharar, Majri, Nawan Shahr, Patiala, Rajpura, Samana, Sanaur, Saroya, Sirhind	Balachaur, Bassi Pathanan, Derabassi, Majri, Rajpura, Sanaur, Sirhind	Bassi Pathanan, Derabassi, Majri, Rajpura
Low-Low	Abohar, Ajnala, Batala, Bhagta Bhai Ka, Fazilka, Lambi, Malout, Raikot, Rayya, Shahkot, Arniwal	Abohar, Batala, Bhagta Bhai Ka, Lambi, Malout, Rayya, Shahkot	Abohar, Bhagta Bhai Ka, Lambi
Low-High	Banga, Bhawanigarh, Ghanaur, Khera, Ludhiana Ii, Morinda, Nabha, Nurpur Bedi, Shambu Kalan	Bhawanigarh, Ghanaur, Khera, Morinda, Nabha, Shambu Kalan	Khera, Morinda
High-Low	Bathinda, Dasuya, Firozpur, Moga I, Tarn Taran, Verka	Bathinda, Dasuya, Firozpur, Moga I, Tarn Taran	Bathinda
November 2016			
High-High	Bassi Pathanan, Bhawanigarh, Bhikhi, Derabassi, Majri, Nabha,	Bassi Pathanan, Bhawanigarh, Bhikhi, Derabassi,	Bhawanigarh, Nabha, Samana

	Patiala, Rajpura, Samana, Sanaur, Sirhind	Nabha, Patiala, Rajpura, Samana, Sanaur, Sirhind	
Low-Low	Abohar, Ajnala, Batala, Bhagta Bhai Ka, Dera Baba Nanak, Faridkot, Fatehgarh Churian, Firozpur, Lambi, Lohian, Makhu, Mamdot, Raikot, Shahkot, Mehtapur	Ajnala, Batala, Dera Baba Nanak, Fatehgarh Churian, Firozpur, Lambi, Makhu, Mamdot, Raikot, Shahkot	Batala, Dera Baba Nanak, Firozpur
Low-High	Bhunga, Budhlada, Ghanaur, Khera, Morinda, Saroya, Shambu Kalan, Dirba	Ghanaur, Khera, Shambu Kalan	Khera, Shambu Kalan
High-Low	Bathinda, Muktsar, Tarn Taran, Verka	Bathinda, Muktsar	Muktsar
December 2016			
High-High	Bassi Pathanan, Budhlada, Derabassi, Nabha, Rampura, Samana, Sanaur, Sirhind	Budhlada, Nabha, Rampura, Samana, Sanaur, Sirhind	NULL
Low-High	Bhawanigarh, Bhikhi, Ghanaur, Jhunir, Khera, Majri, Maur, Morinda, Rajpura, Shambu Kalan	Bhawanigarh, Bhikhi, Ghanaur, Jhunir, Khera, Majri, Maur, Shambu Kalan	Bhikhi
High-Low	Bathinda, Ludhiana I	Bathinda, Ludhiana I	Bathinda, Ludhiana I
August 2017			
High-High	Derabassi, Dhar Kalan, Dhilwan, Kapurthala, Majri, Morinda, Sujanpur, Pathankot	Dhar Kalan, Kapurthala, Majri, Pathankot	Dhar Kalan
Low-Low	Bhagta Bhai Ka, Doraha, Faridkot, Firozpur, Jagraon, Kot Bhai At Gidderbaha, Jaitu, Ludhiana I, Nihal Singh Wala, Raikot, Sidhwan Bet	Faridkot, Jaitu, Ludhiana I	NULL
Low-High	Anandpur Sahib, Jalandhar West, Khera, Nadala, Narot Jaimal Singh, Gharota, Sultanpur Lodhi	Jalandhar West, Narot Jaimal Singh, Gharota, Sultanpur Lodhi	Jalandhar West, Narot Jaimal Singh, Gharota
September 2017			
High-High	Derabassi, Dhilwan, Hoshiarpur Ii, Kapurthala, Majri, Nakodar, Gharota, Sultanpur Lodhi	Dhilwan, Kapurthala, Nakodar, Sultanpur Lodhi	Sultanpur Lodhi
Low-Low	Ajnala, Batala, Bathinda, Bhagta Bhai Ka, Bhikhiwind, Chogawan, Gandiwind Tatla At Chohla Sahib, Khanna, Kot Bhai At Gidderbaha, Kot Ise Khan, Jaitu, Lambi, Ludhiana I, Moga I, Muktsar, Nathana, Sangat, Talwandi Sabo, Tarn Taran, Verka, Attari, Goiniana	Bathinda, Bhagta Bhai Ka, Bhikhiwind, Kot Ise Khan, Lambi, Ludhiana I, Moga I, Nathana, Tarn Taran, Verka	Bathinda, Tarn Taran, Verka

Low-High	Adampur, Bhogpur, Bhunga, Dhar Kalan, Jalandhar East, Jalandhar West, Khera, Morinda, Nadala	Adampur, Bhunga, Jalandhar West	Jalandhar West
October 2017			
High-High	Amloh, Bassi Pathanan, Bhawanigarh, Derabassi, Hoshiarpur Ii, Kharar, Majri, Nabha, Sirhind	Amloh, Bassi Pathanan, Bhawanigarh, Derabassi, Kharar, Majri, Sirhind	Amloh, Sirhind
Low-Low	Abohar, Ajnala, Batala, Bhikhiwind, Chogawan, Chohla Sahib, Dera Baba Nanak, Fatehgarh Churian, Harsha Chhina, Jandiala, Lambi, Majitha, Tarn Taran, Tarsikka, Verka, Attari	Ajnala, Batala, Bhikhiwind, Chogawan, Dera Baba Nanak, Fatehgarh Churian, Harsha Chhina, Jandiala, Lambi, Majitha, Tarn Taran, Tarsikka, Verka, Attari	Ajnala, Batala, Chogawan, Dera Baba Nanak, Jandiala, Majitha, Tarn Taran, Tarsikka, Verka
Low-High	Adampur, Bhunga, Dhar Kalan, Khamanon, Khera, Mamdot, Morinda, Rajpura, Shambu Kalan	Khamanon, Khera, Morinda, Rajpura, Shambu Kalan	Khera, Morinda
High-Low	Bathinda, Dasuya, Gurdaspur, Ludhiana I	Bathinda, Gurdaspur	NULL
November 2017			
High-High	Amloh, Bassi Pathanan, Bhawanigarh, Bhikhi, Kharar, Majri, Nabha, Sirhind, Kot Kapura	Amloh, Bassi Pathanan, Bhawanigarh, Kharar, Nabha, Sirhind	Amloh, Sirhind
Low-Low	Ajnala, Batala, Bhikhiwind, Chogawan, Dasuya, Dera Baba Nanak, Fatehgarh Churian, Jandiala, Majitha, Qadian, Raikot, Shahkot, Tarn Taran, Tarsikka, Verka	Ajnala, Batala, Dasuya, Dera Baba Nanak, Fatehgarh Churian, Majitha, Tarn Taran, Verka	Ajnala, Batala, Dasuya, Fatehgarh Churian, Tarn Taran, Verka
Low-High	Ghanaur, Jhunir, Khamanon, Khera, Maur, Morinda, Rajpura, Samana, Sanaur, Shambu Kalan	Khamanon, Khera, Morinda, Sanaur, Shambu Kalan	Khera, Morinda
December 2017			
High-High	Amloh, Bagha Purana, Bhawanigarh, Kharar, Muktsar, Nabha, Patiala, Samana, Sanaur, Sirhind, Kot Kapura	Amloh, Bagha Purana, Bhawanigarh, Nabha, Samana, Sanaur, Sirhind	Amloh, Sirhind
Low-Low	Ajnala, Batala, Bathinda, Lohian, Mahal Kalan, Majitha, Makhu, Ahemdagarh, Nathana, Patti, Sehna, Shahkot	Batala, Majitha, Ahemdagarh	NULL

Low-High	Bhikhi, Ghanaur, Jhunir, Khamanon, Khera, Morinda, Shambu Kalan	Ghanaur, Khera, Morinda, Shambu Kalan	Khera
High-Low	Fazilka, Nur Mahal, Tarn Taran	Fazilka	Fazilka
August 2018			
High-High	Anandpur Sahib, Andana, Balachaur, Bhawanigarh, Bhikhi, Derabassi, Majri, Nurpur Bedi, Rajpura, Sanaur, Sunam, Dirba	Andana, Bhawanigarh, Bhikhi, Majri, Sanaur, Sunam, Dirba	Bhawanigarh, Bhikhi, Dirba
Low-Low	Ajnala, Batala, Bhagta Bhai Ka, Doraha, Jagraon, Jandiala, Khadur Sahib, Khanna, Kot Ise Khan, Ludhiana I, Majitha, Moga I, Moga Ii, Nihal Singh Wala, Patti, Sehna, Sidhwan Bet, Tarn Taran, Verka, Attari	Jagraon, Khadur Sahib, Kot Ise Khan, Ludhiana I, Majitha, Moga I, Nihal Singh Wala, Sidhwan Bet, Tarn Taran	NULL
Low-High	Barnala, Budhlada, Dhuri, Ghanaur, Jhunir, Khera, Morinda, Samana, Sirhind, Shambu Kalan, Lehra Gaga	Dhuri, Ghanaur, Morinda, Samana, Shambu Kalan, Lehra Gaga	Samana, Shambu Kalan
High-Low	Bathinda	Bathinda	Bathinda
September 2018			
High-High	Bhawanigarh, Bhikhi, Dhuri, Majri, Samana, Sanaur, Sirhind, Shambu Kalan, Dirba	Bhawanigarh, Bhikhi, Dhuri, Samana, Sanaur, Sirhind, Shambu Kalan, Dirba	Bhawanigarh, Bhikhi, Samana
Low-Low	Abohar, Ajnala, Batala, Bhikhiwind, Chogawan, Chohla Sahib, Dasuya, Dera Baba Nanak, Fatehgarh Churian, Fazilka, Firozpur, Gandiwind Tatla At Chohla Sahib, Harsha Chhina, Jagraon, Jandiala, Khadur Sahib, Khanna, Kot Ise Khan, Lohian, Ludhiana I, Mahal Kalan, Majitha, Makhu, Ahemdagarh, Moga I, Moga Ii, Naushera Pannuan, Pakhowal, Patti, Raikot, Rayya, Shahkot, Sidhwan Bet, Sri Hargobindpur, Tarn Taran, Tarsikka, Verka, Attari, Mehtapur	Abohar, Ajnala, Batala, Bhikhiwind, Chohla Sahib, Dasuya, Dera Baba Nanak, Fazilka, Gandiwind Tatla At Chohla Sahib, Jandiala, Khadur Sahib, Kot Ise Khan, Lohian, Ludhiana I, Majitha, Makhu, Ahemdagarh, Moga I, Naushera Pannuan, Patti, Raikot, Rayya, Shahkot, Tarn Taran, Verka, Attari	Bhikhiwind, Chohla Sahib, Dera Baba Nanak, Khadur Sahib, Ludhiana I, Majitha, Moga I, Patti, Rayya, Tarn Taran, Verka
Low-High	Adampur, Anandpur Sahib, Balachaur, Barnala, Budhlada, Ghanaur,	Ghanaur, Nabha, Sunam	Ghanaur

	Hoshiarpur Ii, Jhunir, Mahilpur, Morinda, Nabha, Sunam		
October 2018			
High-High	Balachaur, Barnala, Bhikhi, Dhuri, Faridkot, Rampura, Sanaur, Sirhind	Balachaur, Bhikhi, Dhuri, Faridkot, Sirhind	Bhikhi
Low-Low	Abohar, Ajnala, Batala, Bhikhiwind, Chogawan, Dera Baba Nanak, Fatehgarh Churian, Fazilka, Gandiwind Tatla At Chohla Sahib, Gurdaspur, Kot Ise Khan, Lohian, Ludhiana I, Majitha, Naushera Pannuan, Nur Mahal, Patti, Raikot, Shahkot, Tarn Taran, Verka, Attari, Mehtapur	Ajnala, Batala, Bhikhiwind, Fazilka, Lohian, Ludhiana I, Majitha, Nur Mahal, Raikot, Shahkot, Tarn Taran, Verka, Mehtapur	Ajnala, Fazilka, Ludhiana I, Majitha, Tarn Taran, Verka
Low-High	Anandpur Sahib, Bagha Purana, Bhawanigarh, Budhlada, Ghanaur, Jhunir, Jaitu, Majri, Maur, Morinda, Nabha, Samana, Shambu Kalan, Sunam, Dirba	Anandpur Sahib, Bhawanigarh, Ghanaur, Jaitu, Majri, Maur, Morinda, Nabha, Samana, Shambu Kalan	Bhawanigarh
High-Low	Dasuya, Moga I, Rayya	Dasuya, Moga I, Rayya	Moga I
November 2018			
High-High	Balachaur, Barnala, Dhuri, Faridkot, Jaitu, Muktsar, Rampura, Kot Kapura	Dhuri, Faridkot, Jaitu, Muktsar, Rampura	NULL
Low-Low	Ajnala, Batala, Dasuya, Doraha, Gurdaspur, Ludhiana I, Nur Mahal, Pakhowal, Raikot, Shahkot, Sudhar, Dehlon, Mehtapur	Ajnala, Dasuya, Doraha, Gurdaspur, Ludhiana I, Nur Mahal, Pakhowal, Raikot, Shahkot, Sudhar, Mehtapur	Ludhiana I, Nur Mahal, Mehtapur
Low-High	Anandpur Sahib, Bagha Purana, Bhawanigarh, Bhikhi, Budhlada, Guru Har Sahai, Jhunir, Khera, Kot Bhai At Gidderbaha, Majri, Maur, Morinda, Shambu Kalan, Sunam	Bagha Purana, Bhawanigarh, Bhikhi, Budhlada, Jhunir, Maur	Bhikhi
High-Low	Moga I, Tarn Taran, Verka	Moga I, Tarn Taran, Verka	NULL
December 2018			
High-High	Bhawanigarh, Dhuri, Faridkot, Jaitu, Muktsar, Nabha, Samana, Sanaur, Kot Kapura	Bhawanigarh, Faridkot, Jaitu, Muktsar, Kot Kapura	Faridkot, Jaitu, Muktsar, Kot Kapura
Low-Low	Adampur, Ajnala, Batala, Dasuya, Dera Baba Nanak, Dhariwal, Doraha, Fatehgarh	Ajnala, Batala, Dasuya, Dhariwal, Doraha,	Dasuya, Ludhiana I

	Churian, Fazilka, Gurdaspur, Hoshiarpur I, Hoshiarpur Ii, Ludhiana I, Ahemdagarh, Nadala, Nur Mahal, Pakhowal, Raikot, Rayya, Rurka Kalan, Samrala, Sudhar, Dehlon	Fatehgarh Churian, Ludhiana I, Raikot, Rayya, Rurka Kalan, Samrala	
Low-High	Bagha Purana, Bamial, Bhikhi, Budhlada, Ghanaur, Guru Har Sahai, Jhunir, Kot Bhai At Gidderbaha, Maur, Arniwal, Shambu Kalan, Dirba	Bagha Purana, Bhikhi	Bagha Purana
August 2019			
High-High	Balachaur, Dera Baba Nanak, Dhariwal, Gurdaspur, Kalanaur, Patiala, Shambu Kalan	Dera Baba Nanak, Dhariwal, Gurdaspur, Kalanaur	Dera Baba Nanak, Dhariwal
Low-Low	Jandiala, Khadur Sahib, Kot Ise Khan, Ludhiana I, Nur Mahal, Phagwara, Tarn Taran	Ludhiana I	NULL
Low-High	Banga, Batala, Bhikhi, Dorangala, Khera, Mamdot, Maur, Rajpura, Rampura, Saroya, Talwandi Sabo	Batala, Bhikhi, Mamdot, Rajpura, Saroya	NULL
High-Low	Anandpur Sahib, Kapurthala	Anandpur Sahib, Kapurthala	Anandpur Sahib, Kapurthala
September 2019			
High-High	Amloh, Aur, Balachaur, Dera Baba Nanak, Dhariwal, Kalanaur, Nabha, Sri Hargobindpur	Balachaur, Dera Baba Nanak, Dhariwal, Kalanaur, Nabha	Dera Baba Nanak, Dhariwal, Kalanaur
Low-Low	Abohar, Bhagta Bhai Ka, Lambi, Malout, Nur Mahal, Patran, Raikot, Rurka Kalan, Lehra Gaga	Abohar, Lambi, Nur Mahal	Abohar
Low-High	Dhilwan, Dorangala, Fatehgarh Churian, Garh Shankar, Jalandhar West, Khera, Machhiwara, Majitha, Qadian, Rayya, Saroya, Tarsikka	Fatehgarh Churian, Khera, Qadian, Saroya, Tarsikka	NULL
High-Low	Fazilka	Fazilka	NULL
October 2019			
High-High	Amloh, Aur, Balachaur, Jalandhar East, Kalanaur, Ludhiana Ii, Dehlon	Aur, Balachaur, Kalanaur, Ludhiana Ii	NULL
Low-Low	Abohar, Andana, Bhagta Bhai Ka, Bhunarheri, Chohla Sahib, Dasuya, Ghanaur, Lambi, Patran, Phul, Sanaur, Lehra Gaga	Abohar, Andana, Bhagta Bhai Ka, Bhunarheri, Chohla Sahib, Patran, Lehra Gaga	Lehra Gaga

Low-High	Banga, Dera Baba Nanak, Dhariwal, Dhuri, Fatehgarh Churian, Garh Shankar, Jalandhar West, Khera, Khuian Sarwar, Machhiwara, Majitha, Samrala, Saroya	Dera Baba Nanak, Dhariwal, Machhiwara, Majitha, Saroya	NULL
High-Low	Bathinda, Firozpur, Mansa, Tarn Taran	Bathinda, Firozpur, Tarn Taran	Tarn Taran
November 2019			
High-High	Aur, Balachaur, Chamkaur Sahib, Khera	Balachaur	NULL
Low-Low	Andana, Bamial, Bhagta Bhai Ka, Dasuya, Patran, Raikot, Samana, Shahkot, Sujanpur, Pathankot, Mehtapur, Lehra Gaga	Andana, Bamial, Patran, Samana, Shahkot, Pathankot, Mehtapur, Lehra Gaga	Andana, Patran, Lehra Gaga
Low-High	Anandpur Sahib, Gandiwind Tatla At Chohla Sahib, Jalalabad, Khuian Sarwar, Ludhiana Ii, Machhiwara, Majri, Morinda, Saroya, Arniwal, Attari	Jalalabad, Khuian Sarwar, Arniwal	Khuian Sarwar
December 2019			
High-High	Bassi Pathanan, Jalalabad, Khadur Sahib, Khera, Moga Ii, Naushera Pannuan, Patti, Rajpura	Khera, Naushera Pannuan, Patti	NULL
Low-Low	Ajnala, Doraha, Ludhiana I, Nur Mahal, Patran, Phillaur, Shahkot, Dehlon, Lehra Gaga, Dirba	Ajnala, Doraha, Ludhiana I, Shahkot, Dehlon, Dirba	Ludhiana I
Low-High	Amloh, Anandpur Sahib, Aur, Bagha Purana, Balachaur, Gandiwind Tatla At Chohla Sahib, Kot Bhai At Gidderbaha, Morinda, Sangat, Saroya, Arniwal, Goiniana, Shambu Kalan	Amloh, Balachaur, Goiniana	NULL
High-Low	Batala, Gurdaspur, Mansa, Sudhar	Batala, Sudhar	Batala
Low-Low	Batala, Ludhiana Ii, Rayya	NULL	NULL
High-Low	Arniwal	NULL	NULL
High-Low	Gurdaspur, Hoshiarpur I, Ludhiana I	NULL	NULL
High-Low	Tarn Taran, Verka	NULL	NULL
High-Low	Ajnala, Batala, Bathinda, Firozpur, Moga I	NULL	NULL

APPENDIX G – PLAGIARISM CHECK REPORT

Ouriginal by Turnitin

Document Information

Analyzed document	thesis_plagiarism_check_20230303.docx (D159993598)
Submitted	2023-03-03 13:35:00
Submitted by	Biju Soman
Submitter email	bijusoman@sctimst.ac.in
Similarity	3%
Analysis address	bijusoman.sctims@analysis.urkund.com

Sources included in the report

SA **158)Latest Edited Thesis 07.03.pdf** 7
Document 158)Latest Edited Thesis 07.03.pdf (D130288694)

SA **vikram gupta dissertation.docx** 2
Document vikram gupta dissertation.docx (D30156436)

Entire Document

CHAPTER 1 INTRODUCTION
1 INTRODUCTION 1.1 Research context